Débora Spenassato

# MANUTENÇÃO DO BANCO DE ITENS PARA TESTES ADAPTATIVOS COMPUTADORIZADOS APLICADOS EM AVALIAÇÕES DE ALTO IMPACTO

Tese submetida ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de Santa Catarina para a obtenção do Grau de Doutora em Engenharia de Produção.
Orientador: Prof. Dr. Antonio Cezar Bornia.

Florianópolis
2017

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Spenassato, Débora
MANUTENÇÃO DO BANCO DE ITENS PARA TESTES ADAPTATIVOS COMPUTADORIZADOS APLICADOS EM AVALIAÇÕES DE ALTO IMPACTO / Débora Spenassato ; orientador, Antonio Cezar Bornia - SC, 2017.
252 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós Graduação em Engenharia de Produção, Florianópolis, 2017.

Inclui referências.

1. Engenharia de Produção. 2. Manutenção do Banco de Itens. 3. Testes Adaptativos Informatizados . 4. Teoria da Resposta ao Item. I. Bornia, Antonio Cezar . II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Engenharia de Produção. III. Título.

Débora Spenassato

**MANUTENÇÃO DO BANCO DE ITENS PARA TESTES ADAPTATIVOS COMPUTADORIZADOS APLICADOS EM AVALIAÇÕES DE ALTO IMPACTO**

Esta Tese foi julgada adequada para obtenção do Título de Doutora em Engenharia de Produção, e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de Santa Catarina.

Florianópolis, 22 de fevereiro de 2017.

_____
Prof. Fernando Antônio Forcellini, Dr.
Coordenador do Curso

**Banca Examinadora:**

| | |
|---|---|
| _____<br>Antonio Cezar Bornia, Dr.<br>Orientador<br>EPS/UFSC | _____<br>Dalton Francisco de Andrade, PhD<br>Membro<br>INE/UFSC |
| _____<br>Adriano Ferreti Borgatto, Dr.<br>Membro<br>INE/UFSC | _____<br>Juliano Anderson Pacheco, Dr.<br>Examinador Externo<br>Federação das Indústrias do Estado de Santa Catarina (FIESC) |
| _____<br>Pedro Alberto Barbetta, Dr.<br>Membro<br>INE/UFSC | _____<br>Fernando de Jesus Moreira Jr., Dr.<br>Examinador Externo (via Skype)<br>Universidade Federal de Santa Maria (UFSM) |

Este trabalho é dedicado aos meus queridos pais, Lourdes e Gelson (*in memorian*).

## AGRADECIMENTOS

Agradeço a Deus pelas oportunidades, por me dar força, luz e sabedoria para concluir mais esta tarefa.

Ao meu orientador, professor Antonio Cezar Bornia, pela confiança, amizade, respeito e ensinamentos na construção deste trabalho.

Aos professores Dalton F. de Andrade, Adriano F. Borgatto e Pedro A. Barbetta por compartilharem comigo seus conhecimentos.

Aos meus familiares pela compreensão e apoio.

Aos colegas que passaram pelo Laboratório de Custos e Medidas (LCM - UFSC) durante o período do doutorado, pela amizade, convivência, respeito, cooperação e troca de conhecimentos.

Ao CNPq pelo auxílio financeiro durante este trabalho.

## RESUMO

Teste adaptativo computadorizado (CAT) é um método de aplicação de testes capaz de produzir um teste personalizado para cada indivíduo. O sucesso de um CAT depende, entre outros fatores, de um banco de itens (BI) calibrado pela teoria da resposta ao item (TRI) e com boas qualidades psicométricas. Esse BI necessita de manutenção periódica para não comprometer os resultados da avaliação, principalmente em testes de alto impacto, ou seja, quando seus resultados são usados para tomar decisões importantes que afetam estudantes, professores ou instituições. O objetivo deste trabalho foi desenvolver uma sistemática para a manutenção do banco de itens para testes adaptativos computadorizados aplicados em avaliações de alto impacto. Duas fases foram definidas. A Fase 1 visa acrescentar novos itens no BI (calibração de novos itens), enquanto a Fase 2 visa o monitoramento para identificar itens comprometidos e que precisam ser eliminados do BI (Etapa 1 - monitoramento da exposição dos itens e Etapa 2 - verificação de *drift* dos parâmetros dos itens). Diretrizes são fornecidas para orientar os desenvolvedores de testes e profissionais responsáveis pela manutenção de CATs, sobre como proceder em cada fase de análise. Diferentes simulações são apresentadas para definição do *design* do CAT operacional para avaliar profissionais de um curso de capacitação na área da saúde, bem como os resultados da aplicação do CAT em duas edições de testes e implementação de parte da sistemática de manutenção do BI. Por fim, sugere-se a execução das duas fases da sistemática a cada edição de testes.



**Palavras-chave**: Manutenção do Banco de Itens. Testes Adaptativos Informatizados. Teoria da Resposta ao Item.

## ABSTRACT

Computerized adaptive testing (CAT) is a form of computer-administered test that adapts to the examinee's latent trait level. Adapting the difficulty of items means that an examinee with high latent trait will receive difficult items, while an examinee with low latent trait will receive easy items. The success of a CAT depends, among other factors, on an item bank (BI) with both good psychometric qualities and calibration based on the item response theory (IRT). In order to keep effectiveness, this item bank needs to be periodically up-to-dated so the evaluation results will not be compromised. The item bank maintenance is especially important for high-stakes tests, ie when their results are used to make important decisions that affect students, teachers or institutions. The objective of this work was to develop a systematized approach for the item bank maintenance for computerized adaptive testing applied in high-stakes tests. Two phases have been defined. Phase 1 aims to increase BI (calibration of new items), while Phase 2 aims to identify potentially compromised items that need to be eliminated of the BI (Step 1 - monitoring the items exposure and Step 2 - detection of item parameter drift). Guidelines are provided to guide test developers and professionals responsible for maintaining a CAT, on how to proceed in each phase of the analysis. Different simulations are presented to define the CAT design for the evaluation of health professionals at the end a training course. We also present both the results of the CAT application in two rounds of tests and the implementation of part of the approach for the item bank maintenance. Finally, we suggest the execution of the two phases to each round of tests.



**Keywords**: Maintaining item bank. Computer Adaptive Testing. Item Response Theory.

## LISTA DE FIGURAS

## LISTA DE TABELAS

## LISTA DE QUADROS

## LISTA DE ABREVIATURAS E SIGLAS

*AERA - Annual Meeting of the American Educational Research Association*
AIC - Critério de Informação de Akaike
ANOVA - análise de variância
AP - Parâmetro de aceleração
APA - *Annual Meeting of the American Psychological Association*
ASVAB - *Armed Services Vocational Aptitude Battery*
BC - Restrição de balanceamento de conteúdo
BI - Banco de itens
BIB - Blocos incompletos balanceados
BNI - Banco Nacional de Itens
CAT - Teste adaptativo computadorizado
CCI - Curva característica do item
CUSUM - Procedimentos de soma cumulativa
DIF - Funcionamento Diferencial do Item
DPI - *Drift* dos parâmetros do item
DU - *Design* Universal
EaD - Educação à distância
EAP - Estimação pela média da distribuição *a posteriori*
EB DIF - Abordagem bayes empírico do método Mantel-Haenszel
ECDL - *European Computer Driving Licence*
EP - Erro padrão
ETS - *Educational Testing Service*
EU-EAP - EAP com distribuição *a priori* Beta
EU-MAP - MAP com distribuição *a priori* Beta
FIBI - Função de informação do bando de itens
FII - Função de informação do item
FIPC - Método *fixed item parameter calibration*
FIT - Função de informação de teste
FRI - Função de resposta ao item
GMAT - *Graduate Management Admission Test*
GRE - *Graduate Record Examination*
IACAT - *Annual Conference of the International Association for Computerized Adaptive Testing*
IE - Método da Elegibilidade do item

IIF -Informação Intervalar de Fisher
INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
IRT-LRT - Testes de razão de verossimilhança da TRI
KL - Informação Kullback-Leibler
KLP - Informação Kullback-Leibler Posterior
LM - *Lagrange Multiplier statistic*
LP - Método de programação linear 0-1
MAP - Estimação pela máxima *a posteriori*
MEI - *Maximum expected information*
MEM - Método de máxima verossimilhança marginal com vários ciclos EM
MEPV - *Minimum expected posterior variance*
MFI - Máxima informação de Fisher
MH - Método Mantel-Haenszel
ML3P - Modelo logístico unidimensional de três parâmetros
MLWI - *Maximum likelihood weighted information*
MML - Procedimento de máxima verossimilhança marginal
MPWI - *Maximum posterior weighted information*
MR - Método restrito (*restricted*)
MV - Estimador de máxima verossimilhança
NAEP - *National Assessment of Educational Progress*
NCDIF - *non-compensatory diferential item functioning method*
NCME - *Annual Meeting of the National Council on Measurement in Education*
NCSBN - National Council of State Boards of Nursing
OEM - Estimativa de máxima verossimilhança marginal com um ciclo EM
OIRPI - *Ordered Informative Range Priority Index*
P&P - Teste baseado em papel e lápis
PDI - Índice de densidade proporcional (*Proportional Density Index*)
PG -*progressive method*
PP - *proportional method*
QI - Quociente de inteligência
RAPS - Rede de Atenção Psicossocial
RL - Método de Regressão Logística

$r^{max}$ - taxa máxima de exposição
RMSE - Raiz do erro quadrático médio
RT - Tempo de resposta ao item
SAMU - Serviço de Atendimento Móvel de Urgência
SH - Método Sympson-Hetter
SHGT - Procedimento SH com controle de sobreposição de teste geral
SHT - Procedimento SH com controle de sobreposição de teste
SHTO - Versão on-line do método SHT
SI - índice de adequação (*suitability index*)
SIBTEST - *Simultaneous Item Bias Test*
SLC - método multinomial condicional
STA - *Shadow test*
TAC - Teste Adaptativo Computadorizado
TAI - Teste Adaptativo Informatizado
TBC - Testes Baseados em Computador
TCT - Teoria Clássica dos Testes
TOEFL - *Test of English as a Foreign Language*
TRI - Teoria da Resposta ao Item
UFSC - Universidade Federal de Santa Catarina
UPA - Unidade de Pronto Atendimento
WLE - Estimador ponderado pela verossimilhança

**SUMÁRIO**

# 1. INTRODUÇÃO

## 1.1 CONTEXTUALIZAÇÃO

Instrumentos de medida podem ser utilizados para obter informações sobre traços latentes para auxiliar na tomada de decisão. Traço latente é uma característica que não pode ser mensurada diretamente (ARAÚJO; ANDRADE; BORTOLOTTI, 2009), mas pode ser estimada com base nas respostas dadas pelos indivíduos a um conjunto de itens (teste ou questionário), desenvolvido especificamente para mensurar esta característica, e seus resultados são interpretados em uma escala de medida.

Esses instrumentos são comumente utilizados em diversas áreas do conhecimento, dentre as quais destacam-se psicológica, sociológica ou educacional e tem a finalidade de avaliação, classificação, seleção, monitoramento, diagnóstico, entre outras.

O formato tradicional de aplicação de testes para mensurar um traço latente é por meio de "papel e lápis" (P&P - do termo inglês *paper-and-pencil*); esta forma de avaliação ainda é a mais utilizada. Segundo Moreira Junior (2011), esse método pode demandar elevados custos com papel, tinta de impressão, correção, equipamentos de impressão, espaço físico para armazenamento dos testes, recursos humanos para elaboração, aplicação, fiscalização, transporte e segurança dos testes, além de comprometer a preservação ambiental e sustentabilidade.

Com a evolução da tecnologia, surgiram os Testes Baseados em Computador (TBC), os quais podem ser aplicados da seguinte forma: no formato fixo (linear), ou seja, os itens (ou parte deles) são os mesmos para todos os respondentes como ocorre em testes P&P, porém são administrados via computador; na forma auto-adaptativo, em que as decisões sobre o nível de dificuldade de cada item são feitas pelo respondente antes de cada item ser administrado, ao invés de um algoritmo computacional (VISPOEL, 1998; WISE, 2014); ou ainda, na forma de testes adaptativos, em que um algoritmo seleciona o próximo item a ser administrado com base nas respostas fornecidas pelo indivíduo aos itens anteriormente aplicados.

O teste adaptativo computadorizado (CAT – do inglês *Computerized Adaptive Testing;* TAC ou TAI – Teste Adaptativo Informatizado no português) ganhou destaque por volta de 1980 com a evolução dos próprios computadores (FETZER et al., 2011), apresentando um crescente número de aplicações nas últimas décadas e substituindo os testes tradicionais devido às suas vantagens (LEROUX et al., 2013).

Dentre as principais vantagens, destacam-se: a capacidade de produzir um teste individual para cada respondente; flexibilidade na aplicação; redução no tamanho do teste sem perda de precisão e, consequentemente, da fadiga dos respondentes, que pode comprometer o desempenho dos mesmos; eliminação dos custos com o processo de impressão de testes e sua distribuição; e redução do tempo de divulgação dos resultados.

No entanto, este método necessita do uso de uma metodologia de avaliação mais robusta do que a metodologia tradicional de avaliação, denominada de Teoria Clássica dos Testes (TCT) (HAMBLETON; SWAMINATHAN; ROGERS, 1991) e que permita a comparação entre indivíduos submetidos a testes diferentes e ao longo do tempo. Para isso, a Teoria da Resposta ao Item (TRI) possui um conjunto de modelos com grande potencial e permite a comparação de indivíduos que foram submetidos a diferentes testes, além de trazer benefícios para a análise e interpretação dos resultados e fornecer a base para a seleção dos itens em CAT.

Ao longo dos últimos 20 anos, os CATs têm se tornado cada vez mais importante na avaliação educacional em larga escala (CHANG, 2015). No entanto, pesquisas e aplicações deste método de avaliação ainda são escassas no Brasil.

Em nível mundial, alguns testes que vêm sendo aplicados com sucesso são: *Graduate Record Examination* (GRE), *Graduate Management Admission Test* (GMAT), *National Council of State Boards of Nursing* (NCSBN), *Armed Services Vocational Aptitude Battery* (ASVAB), *Test of English as a Foreign Language* (TOEFL), *European Computer Driving Licence* (ECDL), *National Assessment of Educational Progress* (NAEP), entre outros. No site da *International Association of Computerized Adaptive Testing*

(IACAT, 2015) é possível obter uma lista de testes adaptativos desenvolvidos.

O desenvolvimento de um CAT consiste fundamentalmente de um banco de itens (BI) calibrado pela TRI, ou seja, com os parâmetros dos itens estimados; uma regra para iniciar o teste; método de seleção de itens; método para estimar o traço latente e regra de finalização do teste. Diversas restrições podem ser impostas no algoritmo, dependendo do contexto em que o teste será aplicado.

Em testes de alto impacto (*high stakes tests*), ou seja, quando seus resultados são usados para tomar decisões importantes que afetam estudantes, professores ou instituições (MADAUS, 1988), exige-se rigorosa segurança do teste e do BI para não comprometer os resultados das avaliações (WISE; KINGSBURY, 2000). Por isso, costuma-se utilizar também a restrição de controle da taxa de exposição dos itens, além de outros meios de controle dos usuários.

Uma preocupação recorrente é que, embora possa parecer simples desenvolver um CAT, na prática, esta é uma tarefa complexa que envolve planejamento, implementação e manutenção (WISE, 1997; WISE; KINGSBURY, 2000) e uma equipe composta por profissionais de diferentes áreas. A manutenção está diretamente relacionada ao BI, o qual é um dos principais responsáveis pelo sucesso do CAT.

Um BI computadorizado é responsável por armazenar um conjunto de itens de teste acompanhado de suas classificações e estatísticas, permitindo fácil e rápida recuperação dos dados, bem como facilitar e melhorar a construção de testes (BERGSTROM; GERSHON, 1995). Sendo assim, depende de itens com boas qualidades psicométricas, com atualizações e análises constantes para manter a validade e confiabilidade dos resultados obtidos ao longo do tempo.

De acordo com Chang e Lu (2010), a manutenção é definida como um conjunto de ações necessárias para que um item seja conservado ou restaurado de modo a poder permanecer de acordo com uma condição especificada. No entanto, a correta manutenção de um BI é uma tarefa desafiadora (HAN; GUO, 2011).

## 1.2 DEFINIÇÃO DO PROBLEMA

A importância da manutenção de um CAT, que está diretamente ligada à manutenção de um BI, é destacada por vários autores (BOCK; MURAKI; PFEIFFENBERGER, 1988; STOCKING, 1988a, 1988b, 1994; BERGSTROM; GERSHON, 1995; WISE, 1997; WISE; KINGSBURY, 2000; WANG; KOLEN, 2001; SQUIRES, 2003; WAY, 2006; THISSEN et al., 2007; CHANG; LU, 2010; HAN; GUO, 2011; PASQUALI, 2013).

Esta preocupação com a manutenção de um BI não é à toa, pois gera elevados custos para os responsáveis pela construção e aplicação dos testes no que tange o desenvolvimento de novos itens ou de um BI, o pré-teste de itens e a calibração, além dos gastos com o sistema computacional como um todo (VELDKAMP; MATTEUCCI, 2013).

Procedimentos para uma adequada manutenção do BI devem ser adotados de modo a não desperdiçar esforços econômicos e de pessoal. No entanto, poucos estudos abordam como fazer a manutenção de um BI para CAT de alto impacto de forma ampla, de modo a conter informações (por exemplo, procedimentos disponíveis e sugestões) sobre como proceder em todo o processo de manutenção, o qual envolve tomada de decisão por parte dos responsáveis em várias fases e é uma tarefa contínua enquanto houver aplicações dos testes.

A manutenção de um BI é de extrema importância, principalmente em testes de alto impacto. Estes tipos de testes são comumente utilizados para a contratação, promoção, oportunidades de formação e outros resultados importantes (SQUIRES, 2003). Nesses casos, os transtornos causados por erros sucessivos cometidos ao longo das avaliações, pelo pré-conhecimento dos itens por parte dos respondentes e pela falta de manutenção são ainda maiores, impactando na efetiva avaliação do traço latente que está sendo mensurado e na vida dos respondentes, podendo gerar problemas para as organizações responsáveis pela aplicação dos testes. Por outro lado, itens não podem ser descartados sem necessidade, uma vez geram elevados custos com o seu desenvolvimento.

Neste contexto, os profissionais que buscam informações sobre a manutenção de um BI para CAT, deparam-se com a falta de informações sobre quais procedimentos devem ser adotados e em que momento devem ser aplicados para a efetiva manutenção de um banco de itens. Desta forma, tem-se a seguinte questão de pesquisa: **como manter um banco de itens para testes adaptativos computadorizados aplicados em avaliações de alto impacto ao longo do tempo?**

## 1.3 OBJETIVOS

### 1.3.1 Objetivo geral

Desenvolver uma sistemática para a manutenção do banco de itens para testes adaptativos computadorizados aplicados em avaliações de alto impacto.

### 1.3.2 Objetivos específicos

- Identificar os componentes e os métodos utilizados no desenvolvimento e implantação de CATs;
- Determinar as etapas necessárias para a manutenção de um banco de itens para CATs aplicados em avaliações de alto impacto e a sequência deste processo;
- Definir procedimentos para operacionalizar a avaliação e manutenção de BIs ao longo do tempo.

## 1.4 JUSTIFICATIVA

### 1.4.1 Relevância e contribuição

A importância e necessidade da proposição desta sistemática se dá porque há uma carência de estudos sobre este tema específico, de uma literatura aprofundada direcionada ao contexto de manutenção do BI para CATs aplicados em avaliações de alto impacto, uma vez que um BI envolve tomada de decisões

ao longo do seu processo de desenvolvimento e aplicação (DAVEY, 2011).

Em 1994, Stocking afirmou que a manutenção é uma tarefa que se difere da atividade inicial de desenvolvimento do CAT, sendo este um dos motivos responsáveis pela falta de referencial para orientação. A tarefa de iniciação envolve o desenvolvimento do BI e questões de *design* e implantação do CAT (algoritmos), as quais necessitam de muita atenção dos desenvolvedores de testes. A tarefa de manutenção também deve ser pensada desde o início, pois as regras definidas nesta etapa inicial vão impactar na necessidade de manutenção e uso do BI para o CAT após sucessivas aplicações dos testes.

A sistemática proposta pode ser aplicada, especialmente, em avaliações de alto impacto, com foco em avaliações que visam mensurar um traço latente, desempenho, conhecimento, habilidade ou proficiência dos respondentes. Por exemplo, em avaliações educacionais e testes de seleção. Todavia, nada impede a adequação da sistemática pelo uso de parte dela, conforme a necessidade.

Parte da sistemática pode ser utilizada em áreas que não exigem rigoroso controle de segurança (por exemplo, sem limitar a taxa de exposição dos itens) e manutenção do BI, mas que permanece a exigência de uma escala válida ao longo do tempo para não comprometer os resultados da avaliação. Neste caso, podem-se citar algumas das avaliações psicológicas como de personalidade, avaliações na saúde, qualidade de vida, satisfação dos consumidores, dentre outras.

Torna-se relevante, tanto do ponto de vista teórico quanto prático, o desenvolvimento de uma sistemática que forneça subsídios aos desenvolvedores de CATs, visto que é um tema que está em pleno desenvolvimento, principalmente no Brasil, onde ainda não há aplicações em larga escala desta forma de avaliação.

Do ponto de vista teórico, este referencial estabelece os passos necessários para a manutenção de um BI, apresentando possíveis caminhos a serem seguidos e cuidados que devem ser adotados em cada etapa, os quais são fundamentais para o bom funcionamento dos CATs e para a validade dos resultados. Além

disso, este trabalho apresenta informações sobre os componentes (regras) e os métodos que podem ser utilizados no processo de desenvolvimento de um CAT. Do ponto de vista prático, a aplicação da sistemática proposta auxilia na obtenção de um teste mais seguro, com qualidade, justo para todos os respondentes e passíveis de comparações ao longo do tempo.

Este trabalho contribui para a área de Engenharia de Produção, por exemplo, por possibilitar seu uso em avaliações do processo de ensino-aprendizagem dos estudantes ou, também, na área de gestão de operações, no contexto de "inteligência organizacional", uma vez que auxilia na tomada de decisões para mensuração de traços latentes úteis para o gerenciamento de uma organização. Por exemplo, a sistemática pode ser utilizada por profissionais responsáveis pela avaliação da aprendizagem organizacional, em testes para a seleção de candidatos de uma empresa e para a avaliação de treinamentos ou certificação.

O CAT poderia ser implantado na pós-graduação, por exemplo, como método de aplicação dos testes no processo anual de seleção dos alunos para o acesso ao Mestrado e Doutorado em Engenharia de Produção da UFSC. Uma vez implantado o CAT, a sistemática pode auxiliar os profissionais responsáveis a melhorar a qualidade do BI e mantê-lo atualizado, além de possibilitar a comparação dos traços latentes dos alunos candidatos ou ingressantes entre sucessivas seleções. Atualmente, o teste é aplicado via P&P. Outro exemplo seria a implantação do CAT como forma de aplicação do “teste ABEPRO”, oferecido pela Associação Brasileira de Engenharia de Produção para gerar informação complementar nos processos de admissão de alunos nos programas de pós-graduação, aplicado duas vezes no ano (ABEPRO, 2015).

Esses testes citados são considerados de alto impacto, pois exigem o sigilo dos itens que serão aplicados e seus resultados são utilizados como parte do processo de ingresso na pós-graduação e exigiriam a manutenção do BI. Além disso, a implantação desse método de avaliação poderia trazer muitos benefícios para os alunos e para o programa que desenvolve e corrige esses testes.

Com a popularização dos computadores, a tendência é que os testes passem a ser computadorizados e, também, adaptativos ao respondente, visto que é um método que traz muitos benefícios em relação aos testes tradicionais (não-adaptativos). Além disso, há uma crescente divulgação de pesquisas que vêm utilizando este modo de avaliação com sucesso.

Neste contexto, é importante que os desenvolvedores estejam preparados e saibam efetuar a correta manutenção. Destaca-se, assim, a importância de um referencial que forneça essas diretrizes, no sentido de apresentar ao leitor "o que fazer", "como fazer" e "quando fazer" e destacando a necessidade da avaliação das informações geradas a partir dos procedimentos aplicados para tomar decisões.

### 1.4.2 Ineditismo

Assuntos relacionados à implantação de um CAT, ao desenvolvimento de novos métodos para algoritmos do CAT, à comparação de métodos de seleção de itens e de estimação do traço latente dos respondentes e à avaliação da comparabilidade entre CAT e P&P são amplamente discutidos na literatura (SPENASSATO; BORNIA; TEZZA, 2015), ao contrário do que se observa para a abordagem "manutenção do BI" para CAT de alto impacto, vislumbrando, assim, o ineditismo da pesquisa.

Atualmente, nota-se que o acompanhamento do BI desenvolvido se dá pela fixação da taxa de exposição dos itens e a exclusão dos mesmos após atingirem esta taxa. No entanto, conforme destaca Zhang (2014), nem todos os itens superexpostos são comprometidos, assim como a baixa exposição não garante sua integridade. Por isso, o uso de técnicas isoladas pode se tornar caro e problemático demais, sendo necessário um conjunto de ações para a efetiva manutenção, as quais podem ser esquematizadas, seguindo certa lógica, para manter a qualidade do BI ao longo do tempo.

Para confirmar o ineditismo deste trabalho, realizou-se uma busca de referencial abordando os temas CAT e TRI, em nove bases de dados. Após quatro etapas de refinamento dos artigos

obtidos, 216 artigos foram analisados. Uma busca na Biblioteca Digital Brasileira de Teses e Dissertações sobre o tema CAT resultou em 11 trabalhos, mostrando que estudos nesta área são recentes no Brasil.

A partir dos resultados desta busca de referencial, foi possível diagnosticar algumas características das publicações e confirmar a falta de estudos sobre o tema. Um dos achados é em relação aos objetivos dos estudos, que geralmente abordavam análises comparativas de diferentes métodos para os componentes do algoritmo dos CATs (WANG; VISPOEL, 1998; WANG; HANSON; LAU, 1999; CHEN; ANKENMANN; CHANG, 2000; CHENG; LIOU, 2003; WEISSMAN, 2006; LEROUX et al., 2013), testes para validação de um CAT e comparações dos resultados obtidos via CAT e P&P (STRAETMANS; EGGEN, 1998; WANG; KOLEN, 2001; GWALTNEY; SHIELDS; SHIFFMAN, 2008; ZITNY et al., 2012; RILEY; CARLE, 2012) ou comparação de diferentes testes que foram desenvolvidos para medir o mesmo traço latente, incluindo o teste na forma adaptativa (HUNG et al., 2014; METERKO et al., 2015; PILKONIS et al., 2014).

Um estudo feito por Spenassato, Bornia e Tezza (2015) em 182 artigos, até março de 2014, traz um panorama geral sobre as publicações nesta temática ao analisar características como áreas de aplicação e palavras-chave dos artigos; autores e periódicos com publicações; objetivo, metodologia e sugestões para trabalhos futuros; desenvolvimento de um CAT (assuntos relacionados ao algoritmo); manutenção de um CAT; e *softwares* utilizados.

Constatou-se, também, que as pesquisas que abordam a manutenção do CAT apresentam limitações acerca dos procedimentos necessários para a manutenção do BI para testes de alto impacto. Na maioria dos casos, quando este assunto era discutido nos artigos, aparece apenas como um único parágrafo ou poucos parágrafos no final do trabalho, de forma superficial, destacando apenas a necessidade de controlar a taxa de exposição e de substituir itens superexpostos, sem dar diretrizes completas de como e quando fazer isso, além dos cuidados e análises que devem ser feitas ao longo do tempo.

Notou-se que as análises que deveriam ser aplicadas de forma conjunta para uma adequada manutenção são tratadas isoladamente, ou seja, avaliam apenas o *drift* dos parâmetros dos itens, tratam apenas de métodos para controle da taxa de exposição dos itens, etc (por exemplo, LI, 2008; ABAD et al., 2010; CLARK, 2013; MCLEOD; LEWIS, 1999; VEERKAMP; GLAS, 2000; MEIJER, 2002; VAN DER LINDEN; VAN KRIMPEN-STOOP, 2003; BELOV, 2012; BAN et al., 2001; POMMERICH; SEGALL, 2003; KINGSBURY, 2009). Desta forma, esses documentos não trazem diretrizes necessárias para quem está iniciando este processo de desenvolvimento e manutenção do BI para CAT.

Algumas referências mais completas sobre o tema sendo investigado merecem ser destacadas. Estudos como de Wise (1997), Wise e Kingsbury (2000) discutem vários aspectos importantes para a manutenção de um BI, porém é um trabalho mais direcionado à segurança do teste.

Way (1998) discute técnicas de segurança para proteger o BI contra roubo ou acesso não autorizado, a importância de limitar a exposição dos itens e a sobreposição em testes de alto impacto, bem como alguns problemas associados à manutenção, como a rotação do BI, divulgação de itens e monitoramento do desempenho dos respondentes (detecção de padrões de respostas discrepantes).

Moreira Junior (2011) desenvolveu uma sistemática para a implantação de um CAT devido à carência de estudos que abordavam o princípio da elaboração até a efetiva aplicação. O autor propõe uma sistemática mais ampla, a qual inclui a manutenção de um BI para CAT, porém, aborda apenas a questão de controle da exposição dos itens e a inserção de novos itens no banco, necessitando da inclusão de outros aspectos relevantes para a adequada manutenção e maior detalhamento destas etapas.

Contribuindo com esta temática, Thompson e Weiss (2011) acrescentam que o desenvolvimento de CAT não cessa quando o teste é publicado, mas que são necessárias pesquisas adicionais, verificando se os resultados do CAT operacional coincidem com os resultados esperados com base nas simulações, e fazer a substituição de itens que se tornam superexpostos.

Visando a substituição de itens no BI, Zheng (2014) e Guo (2016) propuseram novos métodos de seleção de itens de pré-teste para calibração em CAT. Zheng (2014) apresentou um trabalho completo sobre a calibração on-line de itens, que aborda métodos de seleção de itens de pré-teste, local de inserção desses itens no teste, método de estimação dos parâmetros dos itens e regra de parada para calibração. Já Guo (2016) apresenta um *design* de (re)calibração on-line com o objetivo de detectar *drift* nos parâmetros dos itens em CAT. No entanto, esses estudos não discutem questões relevantes que levam a exclusão/substituição de itens ou como decidir pela exclusão.

Observou-se que os trabalhos focam em partes importantes do processo de manutenção, mas não abordam de forma completa todas as etapas necessárias para tal. Outras referências relevantes são Bock, Muraki e Pfeiffenberger (1988), Stocking (1988a, 1988b, 1994), Bergstrom e Gershon (1995), Thissen et al. (2007), Han e Guo (2011).

Por fim, apesar dos esforços de estudiosos da área, nota-se uma carência de estudos com diretrizes para a manutenção do BI para CAT de alto impacto, fazendo com que os pesquisadores e desenvolvedores de testes necessitem buscar em várias fontes como deve ser feita a correta manutenção e quais os procedimentos que estão disponíveis na literatura, sendo este o ineditismo apresentado neste trabalho, que consiste na integração de etapas e avaliações para a adequada manutenção do BI, dando suporte para a tomada de decisão por especialistas ao longo do processo de avaliação.

## 1.5 LIMITES DO TRABALHO

### 1.5.1 Quanto ao desenvolvimento e utilização da sistemática

A sistemática é especialmente útil para um tipo específico de CAT - testes de alto impacto com controle da taxa de exposição dos itens e para modelos dicotômicos unidimensionais da TRI. Para desenvolvimento da sistemática, foi assumido que os programas têm um sistema seguro para a aplicação dos testes

quanto ao roubo do BI ou acesso não autorizado e uso de apenas um BI para o CAT operacional e não, o rodízio de vários BIs.

Como são várias as variáveis que podem ser manipuladas em CAT (estratégia de calibração, tamanho da amostra de respondentes, tamanho do banco de itens, modelo de resposta ao item, etc), as quais tem um grande peso sobre os resultados finais e causam diferentes impactos nas estimativas e no uso dos itens do BI, é inadequado sugerir uma regra geral como sendo a melhor de todas, para todas as situações de testes. Por isso, possíveis caminhos e estudos são apresentados, cabendo aos pesquisadores investigar no seu contexto qual é a melhor delas. Assim, os resultados da aplicação do CAT aqui apresentados são particulares do contexto investigado, não podendo ser generalizado para todas as situações reais.

### 1.5.2 Quanto à implementação das etapas de manutenção e aplicação real do CAT

O avaliação utilizada como exemplo neste trabalho tem por objetivo mensurar o conhecimento dos profissionais sobre os temas abordados no curso de capacitação na área da saúde, após a sua conclusão. Este teste não é considerado de alto impacto, pois o mesmo não foi aplicado em um ambiente controlado e seu resultado não era condicionante para obtenção do diploma do curso. Por isso, o BI real, o algoritmo e a aplicação do CAT apresentam várias limitações para a implementação de todo o processo de manutenção e, portanto, somente algumas etapas serão aplicadas na prática.

Devido às peculiaridades do teste, não será implantado a restrição da taxa máxima de exposição dos itens, mas seu impacto será discutido por meio de simulações. Análises do uso dos itens do BI e da exposição dos itens serão fornecidas. A restrição de balanceamento de conteúdo foi considerada no algoritmo, a qual é fundamental para muitos testes.

Não será discutida a comparabilidade entre testes P&P e CAT, uma vez que o teste foi aplicado sempre via computador, obtendo vantagens que são destacadas no trabalho.

## 1.6 ESTRUTURA DO TRABALHO

Para entender o tema abordado neste trabalho, apresenta-se, no Capítulo 2, uma explanação sobre o CAT, as vantagens e desvantagens do seu uso e as etapas para o seu desenvolvimento, que inclui brevemente assuntos relacionados à validade e fidedignidade de testes, o desenvolvimento do BI e a TRI, e a implantação do CAT (principais componentes e *softwares* disponíveis).

No Capítulo 3 a sistemática desenvolvida é apresentada, juntamente com a operacionalização de cada fase. No Capítulo 4, tem-se os procedimentos metodológicos para a revisão de literatura e desenvolvimento da sistemática, e para a definição do *design* e aplicação do CAT em um curso de capacitação na área da saúde.

No Capítulo 5 apresentam-se os resultados dos estudos simulados e da aplicação do CAT em conjunto com algumas etapas da manutenção do BI. Finalmente, a conclusão e sugestões para estudos futuros são apresentados no Capítulo 6.

## 2. TESTES ADAPTATIVOS COMPUTADORIZADOS

Testes adaptativos computadorizados são testes aplicados de forma adaptativa aos indivíduos, cujos itens são selecionados de acordo com o traço latente do respondente, que geralmente é estimado após cada item ser respondido. Nesses testes, os indivíduos podem responder a itens totalmente diferentes e obterem traços latentes comparáveis, pois o algoritmo computacional seleciona os itens de um conjunto de itens já calibrado pela TRI (banco de itens), os quais estão na mesma escala métrica (LUECHT; DE CHAMPLAIN; NUNGESTER, 1998; KOPEC et al., 2008; ZITNY et al., 2012).

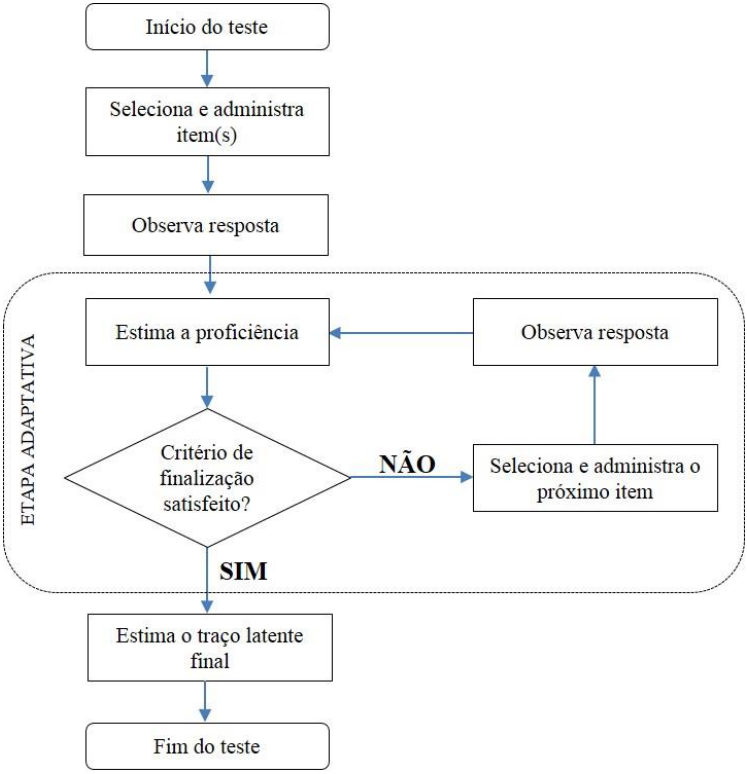
Os primeiros estudos sobre testes adaptativos são datados no início da década de 1900, com os testes de QI (Quociente de inteligência) de Alfred Binet, embora somente nos anos de 1960 e 1970 esses testes começaram a ganhar destaque devido ao avanço da tecnologia nos computadores e investigações no âmbito da TRI; mas, foi a partir de 1980 que muitos pesquisadores implementaram vários tipos de testes adaptativos computadorizados (FETZER et al., 2011). Nas últimas décadas, testes adaptativos têm sido muito utilizados nas áreas de educação, certificação, testes de aptidão, personalidade e na área da saúde.

Os CATs possuem características distintas que variam de teste para teste. No entanto, os componentes básicos (regras) devem ser os mesmos para todas as situações de testes, que compreende um BI calibrado pela TRI, o estabelecimento de um ponto inicial para o teste, método de seleção dos itens, método de estimação do traço latente (ou traço latente do teste) e regra de finalização (parada) do teste. Há várias restrições adicionais que podem ser impostas na seleção dos itens do teste, as principais são: controle da taxa de exposição dos itens e balanceamento de conteúdo.

O algoritmo básico de um CAT é apresentado na Figura 1. O teste inicia com uma estimativa provisória do traço latente do respondente e, em seguida, um item é apresentado com base em algum critério, que geralmente é o item que fornece mais informação para tal nível do traço, observa-se a resposta fornecida,

reestima o traço latente e repete o ciclo até que a regra de parada seja atingida (VISPOEL, 1998; ANATCHKOVA et al., 2012).

Figura 1 – Algoritmo geral de um CAT.



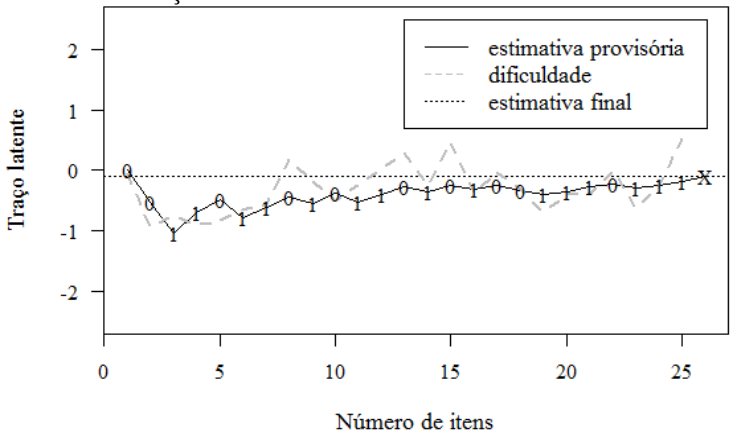

Fonte: Adaptada de Magis e Barrada (2014).

Como os itens são selecionados após a estimação do traço latente, os itens apresentados ao indivíduo acabam sendo mais informativos, e ao se fazer perguntas mais informativas, melhora a eficiência do teste, pois a mesma quantidade de informação é obtida com menos itens e a confiabilidade da escala tende a ser maior se os itens são mais informativos (KOPEC et al., 2008).

Quando um item é administrado e o indivíduo responde corretamente, o próximo item a ser apresentado será mais difícil; caso o indivíduo responda incorretamente, o próximo item será um pouco mais fácil que o anterior, e assim por diante, até atingir a regra de parada do teste. Desta forma, os respondentes serão desafiados, mas não desestimulados por receberem itens muito difíceis ou itens muitos fáceis (WAINER, 2000; CHANG; YING, 2009; DEMARS, 2010).

A Figura 2 apresenta uma ilustração da aplicação do CAT a um respondente com traço latente verdadeiro igual a zero ($\theta$=0). A regra de parada do teste foi fixado em 25 itens e o traço latente passou a ser estimado após a aplicação do primeiro item. É possível observar o padrão de respostas, em que zero significa o erro do item e um significa o acerto. A linha cinza tracejada representa o nível de dificuldade do item. A linha contínua indica a estimativa provisória do traço latente, após cada item aplicado, e a linha pontilhada representa a estimativa final do traço latente ( $\hat{\theta}$ =-0,106 e EP=0,329), o qual está próximo do valor verdadeiro.

Figura 2 – Ilustração do padrão de resposta de um indivíduo submetido a um CAT com 25 itens, nível de dificuldade dos itens apresentados e estimativas do traço latente.




Fonte: Elaborada pela autora.

## 2.1 VANTAGENS E LIMITAÇÕES EM CAT

Os CATs apresentam inúmeras vantagens sobre os testes de P&P e, ainda, reduzem o tempo de teste mantendo a qualidade da mensuração (TIAN et al., 2007; WISE, 1997; WISE; KINGSBURY, 2000). Todavia, algumas limitações deste modo de aplicação de testes também precisam ser destacadas. Dentre os benefícios que podem ser obtidos com a implantação de um CAT, destacam-se:

- Redução de custos com materiais para o desenvolvimento de testes, armazenamento, transporte e correção (WAINER, 2000; MOREIRA JUNIOR, 2011) - o processo de impressão dos testes será eliminado, reduzindo os custos com papel, tinta de impressão e equipamentos. Além disso, reduzem-se custos do processo de correção do teste, o qual é realizado pelo próprio sistema, e na logística de distribuição dos testes, que passam a ser aplicados em computadores;
- Brevidade na divulgação do resultado do teste - o resultado do teste pode ser apresentado imediatamente após o seu término, acompanhado de um relatório que apresenta o traço latente do respondente e a interpretação deste valor. Este *feedback* é mais informativo do que simplesmente oferecer um traço latente final (STRAETMANS; EGGEN, 1998; WAINER, 2000; YI; WANG; BAN, 2001; CELLA et al., 2007);
- Avaliação individualizada - vantajosa para alunos e professores, pois permite conhecer melhor as características de cada respondente, favorecendo para o ensino direcionado às dificuldades dos alunos (OZYURT et al., 2012);
- A dificuldade dos itens é adaptada ao nível do traço latente dos respondentes (LUECHT; DE CHAMPLAIN; NUNGESTER, 1998; VELDKAMP; MATTEUCCI, 2013);
- Economia de itens - redução no comprimento de teste (de aproximadamente 40%) pode ser obtida sem qualquer

perda de precisão da medida (CHANG; YING, 1996; LUECHT; DE CHAMPLAIN; NUNGESTER, 1998; WISE; KINGSBURY, 2000; CELLA et al., 2007; HUANG; LIN; CHENG, 2009; VELDKAMP; MATTEUCCI, 2013);

- Redução no tempo de teste - teste mais curto reduz a fadiga dos respondentes (WISE, 1997; LUECHT; DE CHAMPLAIN; NUNGESTER, 1998; BJORNER et al., 2007; HUANG; LIN; CHENG, 2009);
- Estimativas mais precisas - os testes beneficiam, principalmente, os indivíduos localizados nos extremos de uma escala, onde há pouca informação (CHANG; YING, 1996; CHEN; ANKENMANN; CHANG, 2000; CELLA et al., 2007; HUANG; LIN; CHENG, 2009);
- Modo de apresentação dos itens e facilidade na coleta de respostas - é possível anexar várias multimídias como gráficos, fotografias, áudio ou vídeo e reutilizar o item facilmente (WAINER, 2000; HUANG; LIN; CHENG, 2009; CELLA et al., 2007);
- Aprendizagem e-*learning* - favorece os professores e alunos com uma abordagem mais fácil para a educação à distância - EaD (HUANG; LIN; CHENG, 2009), reduzindo o esforço e tempo necessários à execução das atividades ligadas ao processo avaliativo por parte de professores e tutores (MANSEIRA; MISAGHI, 2013);
- Facilidade na divulgação de mudanças ao longo do tempo (CELLA et al., 2007). Este método permite o monitoramento longitudinal dos resultados, dado que itens diferentes podem ser administrados em momentos diferentes com base no traço latente do respondente;
- Testes sob demanda - é possível porque os respondentes receberão testes diferentes em CAT (STRAETMANS; EGGEN, 1998; WISE; KINGSBURY, 2000; YI; WANG; BAN, 2001). Assim, oferece benefícios práticos como flexibilidade de horários para os respondentes (FINKELMAN; NERING; ROUSSOS, 2009);

- Maior segurança do teste quanto à distribuição e aplicação - diferentemente de testes P&P em que todos os itens que serão aplicados aos respondentes estão disponíveis no teste, em CAT, tem-se um BI com possíveis itens a serem aplicados (TIAN et al., 2007; WAINER, 2000);
- Facilidade para pré-testar itens novos na sequência do teste operacional (WAINER, 2000; TIAN et al., 2007);
- Os CATs tendem a reduzir os efeitos *ceeling* e *floor* (quando os respondentes acertam ou erram todos os itens, não obtendo qualquer informação sobre esse indivíduo), pois selecionam os melhores itens dentre os itens disponíveis no BI (BJORNER et al., 2007; COHEN; SWERDLIK; STURMAN, 2014; KOPEC et al., 2008), beneficiando os indivíduos localizados nos extremos da escala de medida.

Em EaD, CATs podem ser vistos como uma nova opção para a avaliação, tanto para avaliação formativa quanto somativa (HUANG; LIN; CHENG, 2009; OZYURT et al., 2012; KAYA; TAN, 2014) e para tornar o ambiente de aprendizagem mais eficiente com sistemas tutorias inteligentes, adaptando o material didático ao nível de conhecimento do aluno (HUANG; LIN; CHENG, 2009; WAUTERS; DESMET; VAN DEN NOORTGATE, 2010; OZYURT et al., 2012). Após uma investigação, Manseira e Misaghi (2013), constataram que não haviam estudos abrangentes e contínuos de CAT nesses cursos no Brasil.

Algumas limitações ou problemas que podem existir no decorrer do processo de implantação e aplicação de um CAT são:

- Modo de aplicação do teste - a não familiaridade com o computador pode aumentar o estresse e ansiedade influenciando no desempenho do respondente. Por isso, envolve, entre outras coisas, o desenvolvimento de uma *interface* amigável (STRAETMANS; EGGEN, 1998; VISPOEL,1998; KOPEC et al., 2008);
- O BI calibrado pela TRI deve ser suficientemente grande para abranger todos os possíveis valores para o parâmetro

de dificuldade (MOREIRA JUNIOR, 2011; CHANG, 2015; TIAN et al., 2007). Isso é extremamente importante em testes de alto impacto para que a quantidade de itens comuns entre os respondentes seja minimizada e para que os traços latentes sejam estimados com precisão em todos os níveis da escala;

- Riscos à segurança - a flexibilidade de horários para aplicação de testes é uma vantagem, mas, por outro lado, para testes de alto impacto, gera um risco substancial a segurança do teste. Tal flexibilidade possibilita que um respondente se comunique com colegas que irão responder ao teste posteriormente, informando-os sobre quais os tópicos e itens específicos que são aplicados (STRAETMANS; EGGEN, 1998; FINKELMAN; NERING; ROUSSOS, 2009; HUANG; LIN; CHENG, 2009; KAYA; TAN, 2014), podendo comprometer a validade do teste, principalmente em situações em que o BI é pequeno. Existe, também, o risco de invasão do sistema de BI por *hackers*, se este não possuir segurança adequada (WISE; KINGSBURY, 2000);
- Exclusão digital - falta de acessibilidade à internet ou a computadores em alguns lugares (CELLA et al., 2007), principalmente em comunidades carentes. Para aplicação dos CATs, deve haver um computador disponível para cada respondente e eles precisam ter um mínimo de conhecimento em informática (KOPEC et al., 2008; TIAN et al., 2007);
- Custos para desenvolver e manter os CATs - ainda é bastante caro o processo de desenvolvimento e manutenção de CATs; há gastos consideráveis com o desenvolvimento de itens, pré-teste e calibração dos itens, além de gastos com sistema computacional como um todo (VELDKAMP; MATTEUCCI, 2013);
- Os testes com comprimento variável podem ser percebidos como desiguais pelos respondentes em algumas avaliações porque eles receberão testes com comprimentos distintos (TIAN et al., 2007);

- Falta de comparabilidade entre modos de aplicação dos testes (WANG; KOLEN, 2001; ZITNY et al., 2012). A comparabilidade entre os traços latentes de diferentes modos de aplicação é necessária quando eles coexistirem, ou seja, a mesma escala for utilizada para ambos os testes e, também, quando os itens que compõem o BI para CAT forem calibrados por meio de aplicações P&P (WANG; KOLEN, 2001; MILLS; STOCKING, 1995);
- Na maioria dos CATs não é permitida a revisão de itens, pois poderia alterar toda a cadeia de itens administrados. Também, a ação de "pular" itens e depois voltar para responder não é possível (TIAN et al., 2007; WAINER, 2000; WANG; KOLEN, 2001; WISE, 2014).

No estudo de Kopec et al. (2008), percebeu-se que a falta de conhecimentos em informática pode ter criado uma potencial barreira para a participação dos indivíduos na pesquisa feita pelos autores e, para tentar minimizar este problema, alguns indivíduos responderam ao teste por telefone ou pessoalmente na forma de entrevista. Porém, neste estudo na área da saúde, não havia preocupação com a segurança do teste. Para testes de alto impacto, estas formas opcionais de responder ao teste não seriam possíveis, a não ser em casos especiais, para indivíduos com deficiências que necessitam de assistência.

Estudos de aceitação do CAT por parte dos respondentes, realizados na área da saúde, mostraram boa aceitação dos testes; porém, a maioria das dificuldades destacadas está relacionada à *interface* de um CAT, como letra muito pequena na tela do computador, relatório de *feedback* confuso e dificuldade de rolagem das páginas para baixo (FLIEGE et al., 2009; MCDONOUGH et al., 2012).

Para auxiliar nessas questões, existe uma norma internacional responsável por avaliar a qualidade de produto de *software* (ver PITON GONÇALVES, 2013). Além disso, vários guias para padrões de acessibilidade no contexto de tecnologias da informação e eletrônica estão disponíveis na literatura (ver HARNISS et al., 2007; STONE; LAITUSIS; COOK, 2015).

Neste contexto, o conceito de *Design* Universal (DU) tem sido utilizado na elaboração de instrumentos de medida (ver HARNISS et al., 2007; KAMEI-HANNAN, 2008; NUNES et al., 2015; STONE; LAITUSIS; COOK, 2015), o qual engloba o desenvolvimento de produtos (dispositivos, ambientes, sistemas e processos) que possam ser usados pelo maior número de indivíduos possível, independentemente das situações (ambientes, condições e circunstâncias) (HARNISS et al., 2007). Assim, o DU e o CAT podem trazer benefícios para pessoas com deficiências realizarem os testes de forma mais eficiente, reduzindo a fadiga por apresentar itens adequados a cada respondente e obtendo maior precisão na estimativa do traço latente (NUNES et al., 2015).

Em EaD, apesar dos benefícios já citados do uso de CATs, é preciso ter cautela para que os resultados das avaliações realmente representem o traço latente dos respondentes, pois a tecnologia também facilitou a tentativa de trapacear durante o processo de avaliação (KAYA; TAN, 2014).

Nos casos em que a avaliação deve ser controlada, sugere-se (KAYA; TAN, 2014): a) limitar o tempo para fazer o teste, pois com pouco tempo disponível, há menos oportunidade para o respondente procurar informações; b) exigir que os respondentes façam os testes em locais supervisionados; c) exigir que os respondentes usem uma câmera em seu computador, entre outros. Caso o teste não exija rigorosa segurança, os indivíduos podem responder ao teste em casa ou em qualquer lugar que tenha acesso à internet, como ocorre, por exemplo, em testes de diagnóstico na área da saúde e testes para avaliar satisfação.

## 2.2 DESENVOLVIMENTO DO CAT

As etapas a serem seguidas para obter sucesso no desenvolvimento de um CAT são fontes de várias pesquisas. No entanto, até meados de 2011, poucos estudos fornecendo orientações práticas sobre o desenvolvimento de um CAT estavam disponíveis (THOMPSON; WEISS, 2011; MOREIRA JUNIOR, 2011).

Algumas especificações do teste precisam ser formuladas desde o início de sua viabilização, o que exige certo tempo e recursos para se obter um *design* completo do teste (VELDKAMP; MATTEUCCI, 2013). No contexto geral de CATs, a definição dos componentes e o método adotado em cada um deles dependem de fatores como o tipo de teste, objetivo do teste, tamanho e qualidade do BI, modelo de resposta ao item utilizado, público alvo para a aplicação do teste, entre outros (MOREIRA JUNIOR, 2011).

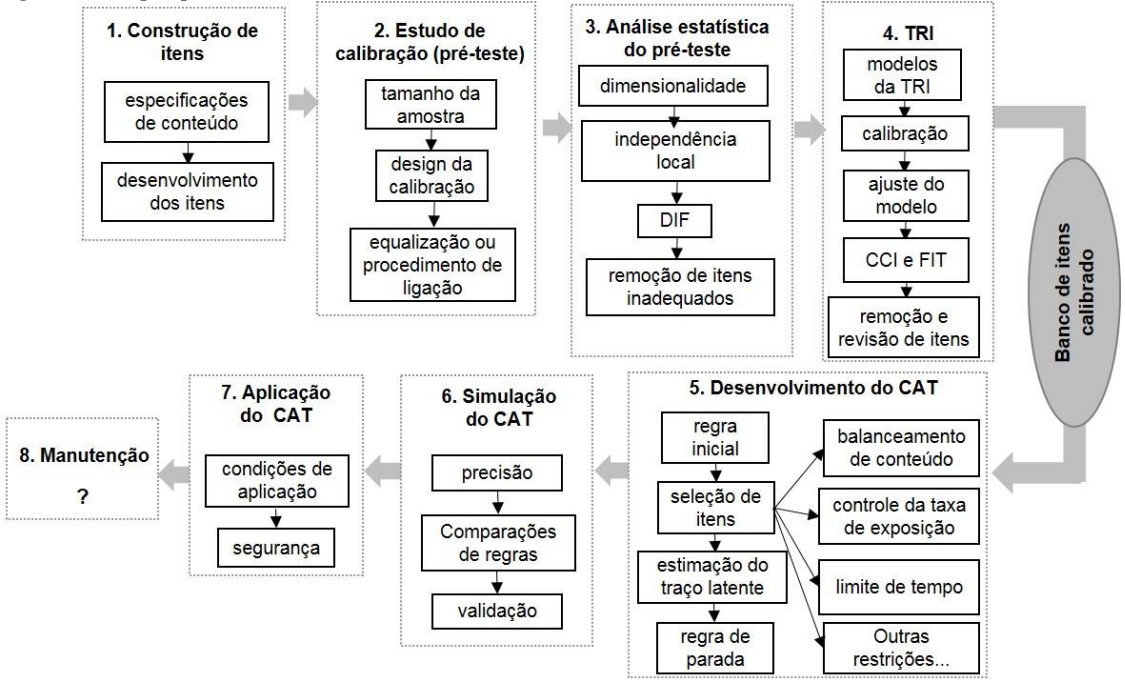
O desenvolvimento de um BI é uma etapa importante no processo de implantação de um CAT, sendo esta, a etapa inicial deste processo (ver WALKER et al., 2010; CELLA et al., 2007; MOREIRA JUNIOR, 2011; THOMPSON; WEISS, 2011; WISE; KINGSBURY, 2000). Após o desenvolvimento do BI, tem-se as etapas relacionadas ao algoritmo e aplicação do teste, conforme representado na Figura 3. Essas etapas serão discutidas ao longo deste trabalho, com atenção especial para a manutenção do BI.

Mais informações sobre orientações para desenvolvimento de um CAT e do BI podem ser obtidas em estudos de Green et al. (1984), Stocking (1994), Ward e Murray-Ward (1994), Bjorner et al. (2007), Thissen et al. (2007), Davey (2011), Lai et al. (2011), McDonough et al. (2012) e Anatchkova et al. (2012).

### 2.2.1 Validade e fidedignidade do teste

A avaliação da qualidade dos testes tem como aspectos fundamentais a validade e a fidedignidade do resultado dos testes (URBINA, 2007; COHEN; SWERDLIK; STURMAN, 2014). Essas evidências dão suporte para a interpretação legítima do que está sendo avaliado e quão bem isso é feito por meio das respostas coletadas em um instrumento (PRIMI; MUNIZ; NUNES, 2016).

Figura 3 – Etapas para o desenvolvimento de um CAT.




Fonte: Adaptada de Spenassato, Bornia e Tezza (2015)

Definições para a validade de um teste têm passado por mudanças em sua definição ao longo do tempo. No entanto, a validade de construto, validade de conteúdo e validade de critério são citadas com maior frequência (PASQUALI, 2013; COHEN; SWERDLIK; STURMAN, 2014). No contexto de testes adaptativos, essas condições também precisam ser verificadas no desenvolvimento do BI e por meio de simulações do CAT.

Devido às constantes discussões e diferentes interpretações do conceito de validade por pesquisadores, a *American Psychological Association* (APA) estabeleceu os *Standards for Educational and Psychological Testing*, desenvolvido em conjunto com a *American Educational Research Association* (AERA) e a *National Council on Measurement in Education* (NCME) desde 1966, que visa assegurar que os testes apresentem os parâmetros de qualidade cientificamente exigidos. Em 2014, uma nova versão atualizada foi lançada.

Conforme Pasquali (2007, p. 105), em seu conceito original, "validade do instrumento diz respeito exclusivamente à pertinência do instrumento com respeito ao objeto que se quer medir; é a questão da referência", sendo o construto, o referencial para os resultados de um teste. Além disso, o autor define construto da seguinte forma:

> O construto (traço latente, teta) se posiciona como o objeto que o teste quer medir, isto é, ele é aquilo que o teste pretende medir. Então, ele é o referente, em função do qual a qualidade do teste deve ser avaliada. Consequentemente, as respostas ao teste (o escore no teste, o observável, o *tau*) não criam o construto, antes, pelo contrário, é o escore no teste que depende do construto (PASQUALI, 2007, p. 106).

"Um teste tem validade de conteúdo se ele constitui uma amostra representativa de um universo finito de comportamentos (domínio)" (PASQUALI, 2013, p. 188). Para viabilizar um teste

com validade de conteúdo, especificações do teste devem ser feitas por especialistas antes da construção dos itens.

Essas especificações abordam (PASQUALI, 2013): 1) definição do conteúdo; 2) explicitação dos objetivos ou processos a serem avaliados; e 3) determinação da proporção no teste de cada tópico do conteúdo. Posteriormente, tem-se as etapas de construção do teste (elaboração dos itens), análise teórica dos itens (análise semântica e de juízes) e análise empírica dos itens (após a aplicação do teste) pela TRI. Uma vez finalizada essas etapas, se espera ter um conjunto de itens representativos do traço latente que se pretende medir.

Esse tipo de validade é frequentemente utilizado no contexto educacional, no qual é usual verificar se os itens que compõem um teste para mensurar o desempenho em uma área específica de conhecimento cobrem todos os conteúdos englobados por tal área ou se é proporcional ao material contemplado no curso (COHEN; SWERDLIK; STURMAN, 2014; PRIMI; MUNIZ; NUNES, 2016).

No CAT, como são selecionados os itens mais informativos para o respondente, nada garante que serão administrados itens de todos os domínios. Em alguns casos, esta condição deve ser garantida para que o traço latente represente adequadamente o construto avaliado. Por isso, deve-se implementar a restrição de balanceamento de conteúdo no processo de seleção de itens de um CAT (HUFF; SURECI, 2000). Esta restrição é apresentada na seção 2.2.3.3.1.

A fidedignidade ou confiabilidade de um teste está relacionada à capacidade de um teste de medir sem erros (PASQUALI, 2013). No contexto de testes e medidas, "a fidedignidade se baseia na consistência e precisão dos resultados do processo de mensuração" (URBINA, 2007, p. 111).

Cohen, Swerdlik e Sturman (2014) e Urbina (2007), reiteram que o erro de mensuração diz respeito a qualquer flutuação nos traços latentes causados por fatores relacionados ao processo de mensuração que são irrelevantes ao que está sendo medido, influenciando no uso específico do teste. Este erro, quando sistemático e consistente, afeta a fidedignidade e a validade

dos resultados (URBINA, 2007). Conforme Pasquali (2013), a variabilidade da amostra e o comprimento do teste são fatores que afetam a fidedignidade do teste.

Vários tipos de evidências estatísticas de confiabilidade podem ser fornecidos, incluindo coeficientes de confiabilidade e generalização, funções de informação, erros padrão de medida, erros padrão condicionais de medida, intervalos de confiança, entre outros (ETS, 2014/2015; URBINA, 2007). Essas evidências devem ser apropriadas para os modelos psicométricos utilizados e para o uso pretendido do traço latente (ETS, 2014/2015).

Para um teste adaptativo, podem-se fornecer estimativas de confiabilidade que levam em conta os efeitos de diferentes métodos na seleção de itens por meio de estudos de simulação do CAT (MOREIRA JUNIOR, 2011; ETS, 2014/2015).

#### 2.2.1.1 Imparcialidade e Funcionamento Diferencial do Item

A imparcialidade (justiça) do teste refere-se ao grau em que um teste é usado de forma justa, equitativa e imparcial (COHEN; SWERDLIK; STURMAN, 2014). Sendo assim, os aspectos de equidade são de responsabilidade daqueles que desenvolvem, usam e interpretam os testes (AERA, APA, NCME, 1999). Por isso, cuidados devem ser tomados antes, durante e após as aplicações dos testes para assegurar testes justos e válidos (SOLÓRZANO; 2008), e que permitam a comparação entre os respondentes.

A definição mais útil de justiça para os desenvolvedores de teste é dada pela medida em que as inferências feitas com base em resultados de testes são válidas para diferentes grupos de respondentes (ETS, 2014/2015; SOLÓRZANO; 2008). A ideia de igualdade de oportunidades ou igual acesso a recursos ou informações que ajudarão os respondentes a obterem melhores desempenhos em avaliações está relacionada ao conceito de justiça (SOLÓRZANO; 2008).

Viés de teste, no contexto dos psicometristas, refere-se a um fator inerente em um teste que sistematicamente impede mensurações corretas e imparciais (COHEN; SWERDLIK;

STURMAN, 2014). A presença de viés em um teste causa o Funcionamento Diferencial do Item (DIF), tornando o processo de avaliação injusto (ANDRIOLA, 2006), impactando na validade da interpretação dos resultados do teste.

Um item é considerado enviesado ou com DIF se indivíduos com o mesmo traço latente possuem diferentes probabilidades de resposta ao item pelo fato de pertencerem a grupos distintos (HOLLAND; WAINER, 1993; ANDRIOLA, 2001, 2006; MAKRANSKY; GLAS, 2013). Isso indica que outras variáveis latentes estão provavelmente influenciando na resposta, cujo comportamento é manifestado nos parâmetros dos itens.

Portanto, um item que exibe DIF tem diferentes parâmetros dos itens, dependendo do grupo ao qual o indivíduo pertencer (EDWARDS, 2009). Porém, pode ocorrer uma mudança nos parâmetros dos itens em testes subsequentes, quando comparado ao seu valor original, denominado *drift* dos parâmetros do item - DPI (WELLS; SUBKOVIAK; SERLIN, 2002).

Neste contexto, técnicas estão disponíveis para prevenir ou reparar esses impactos adversos, como a eliminação de itens com base no DIF ou que não se ajustam ao modelo utilizado (COHEN; SWERDLIK; STURMAN, 2014; ETS, 2014/2015; HUFF; SIRECI, 2000) ou mantê-los no BI, mas com parâmetros diferentes entre os grupos (BJORNER et al., 2007; THISSEN et al. 2007; HART et al., 2009). Neste caso, a informação referente a quem o item pode ou não administrado, pode ser obtida por meio do cadastro do respondente, antes dele iniciar o teste. Por exemplo, quando um item apresentar comportamento diferenciado para os gêneros masculino e feminino.

A presença do DIF em itens de instrumentos de medida é um grave problema que atenta contra o pressuposto da padronização ou uniformização das condições de avaliação; é uma fonte de injustiça, já que produz falta de equidade nos processos avaliativos, favorecendo ou prejudicando o rendimento de um grupo sobre o outro (ANDRIOLA, 2001, 2006).

A análise de DIF pode ser aplicada sob diferentes contextos, incluindo cultura, nacionalidade, regiões geográficas, linguagem, nível educacional, tempo de resposta, idade, gênero, raça, etnia,

diferentes características na área saúde (parte do corpo afetada, histórico cirúrgico, frequência dos sintomas, tempo de lesão), entre outros. O ideal é a elaboração de testes que não contenham itens que apresentem DIF.

O DIF pode ser observado sob duas formas: (1) **DIF uniforme** - está relacionado à dificuldade do item e significa que a diferença na probabilidade de resposta ao item é constante em todos os níveis do traço latente dos respondentes e (2) **DIF não-uniforme** - está relacionado ao parâmetro de discriminação do item e significa que a diferença na resposta ao item vai mudar em diferentes níveis do traço latente, ou seja, não ocorre igualmente em todos os pontos no traço latente, mas são mais evidentes em níveis elevados ou baixos da escala (JETTE et al., 2008; MAKRANSKY; GLAS, 2013).

Hart et al. (2009) alertam que alguns *softwares* para análise do DIF são dependentes do tamanho da amostra, principalmente para DIF não-uniforme, portanto, é importante utilizar amostras equilibradas nos dois grupos para evitar falsos-positivos ou falsos-negativos. Jette et al. (2008) afirmam que, em amostras grandes, o DIF tende a ser significante.

Para análise do DIF, é comum a classificação dos respondentes em dois grupos: o grupo focal - de interesse primário e o grupo de referência - aquele que será comparado com o focal. Para qualquer processo de análise DIF, é preciso selecionar respondentes nos dois grupos correspondentes ao traço latente, antes de comparar o seu desempenho no item avaliado para DIF (NANDAKUMAR; ROUSSOS, 2004).

Foram propostos muitos métodos para avaliar a existência de DIF em itens, sua magnitude e significância. Esses métodos são classificados como não-paramétricos e paramétricos segundo Holland e Wainer (1993). Os métodos mais utilizados são:

- Mantel-Haenszel (MH) (MANTEL; HAENSZEL, 1959) e suas variações (HOLLAND; THAYER, 1988);
- *Simultaneous Item Bias Test* (SIBTEST) (SHEALY; STOUT, 1993) - Funcionamento diferencial de itens e do teste;
- Regressão Logística (RL) (ROGERS; SWAMINATHAN, 1989), sendo este um dos métodos mais utilizados;

- Procedimentos de padronização (*Standardization*) (DORANS; KULICK, 1986);
- Métodos baseados na TRI: testes de razão de verossimilhança da TRI (IRT-LRT) (THISSEN; STEINBERG; WAINER, 1988), teste de Lord (1980), teste de Raju (1988) e análise log-linear (THISSEN; MOONEY, 1989; KELDERMAN, 1990).

De acordo com Lei, Chen e Yu (2006), os testes MH e SIBTEST são mais poderosos na detecção de DIF uniforme do que na detecção de DIF não-uniforme; já RL e IRT-LRT podem detectar os dois tipos de DIF. Um método muito utilizado na TRI é a análise do DIF por meio da comparação da Curva Característica dos Itens - CCIs.

Neste caso, o mesmo item é plotado separadamente para cada grupo que o pesquisador deseja avaliar. Se as CCIs forem idênticas para cada grupo, ou muito parecidas, pode-se dizer que o item não apresenta DIF; se as CCIs são significativamente diferentes, diz-se que apresentam DIF (COHEN; SWERDLIK; STURMAN, 2014).

Itens com DIF são ainda mais problemáticos em CAT, pois diferentes itens são aplicados aos respondentes e não há como saber previamente se muitos itens com DIF serão destinados ao mesmo respondente ou a um grupo, podendo prejudicar seu desempenho ou produzir uma estimativa tendenciosa do traço latente (HART et al., 2009; MAKRANSKY; GLAS, 2013).

### 2.2.2 Banco de itens para CAT

Termos em inglês comumente utilizados para banco de itens são "*item banks*" ou "*item pools*". Bergstrom e Gershon (1995) destacam que há uma distinção entre esses termos: *item banks* refere-se a um conjunto de itens calibrados pela TRI e colocados em uma escala comum; já *item pools* refere-se a um conjunto de itens agrupados por conteúdo, mas não calibrados. Segundo os autores, essa distinção não foi adotada e, atualmente, ambos os termos são usados alternadamente.

Um banco de itens (BI) refere-se ao armazenamento dos itens e de suas características (BERGSTROM; GERSHON, 1995; WALKER et al., 2010; VELDKAMP; MATTEUCCI, 2013), incluindo um controle do fluxo desde sua elaboração até sua aplicação. Para Andrade, Tavares e Valle (2000, p. 85), "bancos são formados por conjuntos de itens que já foram testados e calibrados a partir de um número significativo de indivíduos de uma dada população".

O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) define o banco nacional de itens (BNI) como "uma coleção de itens de testes de natureza específica - organizada segundo determinados critérios - disponíveis para a construção de instrumentos de avaliação. A manutenção do BNI depende da entrada constante de itens de qualidade" (INEP, 2010, p. 5).

A maioria dos desenvolvedores de CATs utilizam a TRI no processo de implementação do teste por apresentar muitas vantagens em relação aos outros métodos, além de fornecer as informações necessárias para a seleção de itens. Uma grande vantagem é que ela permite que diferentes itens sejam administrados a diferentes respondentes e que os traços latentes de todos os respondentes estejam na mesma escala, permitindo, assim, a comparação entre eles (THISSEN et al., 2007; BJORNER et al., 2007; WALKER et al., 2010).

Para possibilitar essa comparação, uma informação fundamental é a escala em que os itens foram calibrados, pois comparações diretas de traços latentes só poderão ser feitas entre indivíduos que tenham seus traços latentes estimados nesta mesma métrica (ANDRADE; TAVARES; VALLE, 2000). Esta escala é arbitrária e o importante são as relações de ordem existentes entre seus pontos; logo, ela é definida por uma média e desvio padrão de alguma população de referência (ANDRADE; TAVARES; VALLE, 2000). Usualmente, utiliza-se a distribuição normal com média zero e desvio padrão um, N(0,1).

A fim de verificar as vantagens da utilização da TRI em CATs, Schnipke e Green (1995) compararam o seu uso (maximizando a informação pela TRI) com outro método de

seleção de itens, o qual busca maximizar a diferenciação (sem usar a TRI) entre grupos de respondentes. Como tais métodos levam a diferentes seleções de itens, os resultados mostraram que os testes baseados na TRI são superiores, especialmente quando o comprimento do teste aumenta e quando um item precisa ser substituído.

Zitny et al. (2012) também destacaram que os algoritmos CAT baseados na TRI para seleção de itens e para estimar o traço latente oferecem oportunidades atraentes para, simultaneamente, otimizar a precisão e eficiência do teste. Além disso, economia de itens são maiores em CAT com uso da TRI (RUDICK; YAM; SIMMS, 2013).

Conforme Ozyurt et al. (2012), a qualidade dos itens do BI determina a qualidade dos CATs. Portanto, deve possuir, entre outras coisas, validade de conteúdo (cobertura de todos os aspectos do construto a ser medido) e ter itens suficientes para obter alta precisão de medida em todos os níveis da escala (BJORNER et al., 2007). Ou seja, itens com vários níveis de dificuldade, permitindo que o algoritmo selecione itens com nível de dificuldade próximo ao nível do traço latente estimado de cada respondente.

O nível de precisão aceitável para determinados níveis da escala pode variar de acordo com o objetivo do teste. Segundo Bjorner et al. (2007), na área da saúde, por exemplo, uma avaliação usada em um ensaio clínico, seria provável demandar alta precisão de medição em todos os níveis da escala para evitar efeitos *ceiling* e *floor*; já para uma avaliação usada no tratamento clínico de pacientes, exige-se alta precisão em níveis baixos de saúde, para os quais é necessário tratamento e acompanhamento, mas não para bons níveis de saúde, em que a intervenção e acompanhamento são desnecessários.

Alta precisão em todos os níveis da escala deve ser exigida em avaliações de alto impacto, dado que pequena variação nos resultados pode, por exemplo, resultar no acesso ou não ao ensino superior ou na contratação ou não de um empregado. Porém, itens localizados nos extremos de uma escala são difíceis de serem avaliados, dado que, quando se utilizam indicadores da psicometria clássica, como correlação ponto bisserial, itens localizados nos

extremos tendem a ser excluídos dos testes por apresentarem um desempenho pobre (BJORNER et al., 2007).

Uma questão importante para um BI é a segurança, principalmente em testes de alto impacto. Em testes educacionais, por exemplo, a avaliação deve ser realizada em um ambiente controlado e itens precisam ser mantidos em sigilo para evitar fraudes (BJORNER et al., 2007). Quando a segurança dos testes é comprometida, prejudica a validade das inferências baseadas no traço latente, não importando o quão forte são as características psicométricas do teste (WISE; KINGSBURY, 2000).

As organizações de testes devem ter segurança apropriada em função da vulnerabilidade de sites e servidores que hospedam as informações de um CAT. Por isso, Way (2006) destaca que o sistema que fornece um CAT deve ser seguro em múltiplos níveis, como: a segurança dos itens do teste deve ser mantida por meio de rotinas de criptografia e a transmissão de dados entre um servidor central e os computadores deve ser monitorada e registrada; o acesso ao sistema nas escolas ou organizações onde serão aplicados os testes deve ser seguro e protegido por senha; as salas de aplicações dos testes devem ser supervisionadas e o acesso ao conteúdo só poderá ser efetuado por pessoal autorizado, além de ser proibido o acesso à internet ou outro ambiente de ferramentas de trabalho.

Para obter maior segurança dos itens e do teste, um grande BI é necessário. A literatura aponta vários **tamanho ideal do BI** para CAT, os quais geralmente variam de 5 a 10 vezes o número de itens a serem aplicados para os respondentes (ver STOCKING, 1994; SEGALL, 2005; DAVEY, 2011; OZYURT et al., 2012; WAY, 1998).

De acordo com Wise e Kingsbury (2000), os fatores que podem influenciar no tamanho do BI são a quantidade de restrições impostas no algoritmo de seleção de itens (quanto mais restrições, maior a quantidade de itens no BI) e a quantidade de aplicações do CAT, principalmente em testes de alto impacto, em que os itens são sigilosos. Nesses casos, o BI precisa ser grande.

Conforme Stocking (1994), o tamanho do BI também depende do critério de finalização adotado para o teste e dos

requisitos de paralelismo com outras formas de testes existentes, como P&P. Para Davey (2011), na prática o tamanho do BI depende de uma variedade de fatores, que acaba sendo determina pelo peso entre os benefícios de BIs maiores (maior eficiência e mais segurança) contra os custos práticos e financeiros do desenvolvimento e pré-teste de um número maior de itens.

Para pré-testar um grande BI, um *design* de equalização ou ligação pode ser utilizado para que os indivíduos respondam a um subconjunto de itens (VELDKAMP; MATTEUCCI, 2013) e o teste não se torne muito longo. Esse *design* também serve para tornar o traço latente dos respondentes comparáveis. Estudos de Ariel, van der Linden e Veldkamp (2006), Veldkamp e van der Linden (2000) e Belov e Armstrong (2009) abordam como construir um BI com características ideais para CAT.

#### 2.2.2.1 Teoria da resposta ao item

Esta metodologia tem como característica a especificação de uma função matemática que representa a relação entre a probabilidade de um indivíduo dar certa resposta a um item como função dos parâmetros do item e do traço latente do respondente (ANDRADE; TAVARES; VALLE, 2000), tendo como foco central o estudo individualizado das características dos itens de um teste (ANDRIOLA, 2001).

Pela TRI, as características dos itens são independentes do grupo utilizado para obtê-las e a estimativa do traço latente do respondente não depende do teste (invariância). Assim, os itens podem ser aplicados novamente a outro grupo (BAKER, 2001; WISE, 1997). Outro ponto importante é que o modelo não requer testes paralelos para avaliar a confiabilidade (HAMBLETON; SWAMINATHAN; ROGERS, 1991; DE AYALA, 2009).

A TRI permite a avaliação da precisão para cada nível em particular, permitindo identificar a faixa do traço latente para o qual o item pode discriminar melhor entre os indivíduos e revelar quão bem diferentes itens discriminam em diferentes níveis (FLIEGE et al., 2009). Esta informação pode ser usada para selecionar os itens mais informativos (discriminativos) e

administrá-los ao respondente em CATs. Assim, utilizar a TRI em conjunto com o CAT, faz com que a confiabilidade adequada seja obtida com erro de mensuração mínimo em testes mais curtos do que no método tradicional (URBINA, 2007).

A TRI dispõe de vários modelos, os quais dependem fundamentalmente de três fatores: (1) da natureza do item (dicotômicos ou politômicos); (2) do número de populações envolvidas; e (3) da quantidade de traços latentes que está sendo medida (unidimensionais ou multidimensionais) (ANDRADE, TAVARES; VALLE, 2000). Os modelos podem ser acumulativos ou não acumulativos (ARAÚJO; ANDRADE; BORTOLOTTI, 2009).

Em uma revisão bibliográfica feita por Spenassato, Bornia e Tezza (2015), concluiu-se que o modelo logístico unidimensional de três parâmetros (ML3P - BIRNBAUM, 1968) foi o mais utilizado em CAT. Como existem vários modelos disponíveis, cabe ao pesquisador decidir qual deles é mais adequado para sua pesquisa.

Testes de alto impacto costumam apresentar itens com múltiplas opções de respostas, os quais são corrigidos como corretos ou incorretos. Devido à estas características, geralmente, modelos da TRI para respostas dicotômicas são utilizados, em especial, o ML3P.

Neste modelo, a probabilidade $p_{ij}$ de que o indivíduo $j$ responda corretamente ao item $i$, é dada pela função de resposta ao item (FRI) em (1):

$$p_{ij} = P\left(U_i = 1 \mid \theta_j\right) = c_i + \left(1 - c_i\right)\frac{e^{a_i\left(\theta_j - b_i\right)}}{1 + e^{a_i\left(\theta_j - b_i\right)}} \qquad (1),$$

onde $i = 1, ..., I;$ e $\theta$ é o traço latente do indivíduo $j$; $P\left(U_i = 1 \mid \theta_j\right)$ é chamada de FRI ou curva característica do item (CCI). O parâmetro $b_i$ representa a dificuldade do item $i$, medido na mesma escala do traço latente, $b_i \in \left(-\infty, +\infty\right)$. O parâmetro $a_i > 0$ é proporcional à inclinação da CCI no ponto $b_i$ e corresponde a discriminação do item $i$. Quanto maior o valor deste parâmetro,

maior será o poder do item em discriminar os respondentes. O parâmetro $c_i$ representa a probabilidade de indivíduos com baixos níveis do traço latente responderem corretamente o item *i;* $c_i \in [0,1]$.

**Pressupostos da TRI**

O ML3P baseia-se no fato de que os indivíduos com maior traço latente possuem maior probabilidade de acertar o item e que esta relação é não-linear (ANDRADE; TAVARES; VALLE, 2000). Este modelo pressupõe algumas limitações, como monotonicidade, unidimensionalidade e independência local. Quando estes pressupostos são violados, sérias consequências podem ocorrer no processo de estimação.

A suposição de monotonicidade significa que a probabilidade de responder corretamente o item deverá aumentar à medida que o traço latente aumenta (REEVE et al., 2007; ARAÚJO; ANDRADE; BORTOLOTTI, 2009). Pode-se dizer, então, que a FRI é estritamente não-decrescente.

Para avaliar a adequação do modelo escolhido aos dados empíricos, utilizam-se estatísticas de ajuste. Há muitos procedimentos disponíveis na literatura, podem-se citar: $\chi^2$ e análise dos resíduos (PASQUALI, 2013); para avaliar a monotonicidade, Reeve et al. (2007) e Smits, Cuijpers e van Straten (2011) sugerem o uso da escala de Mokken, a qual é baseada num modelo da TRI não-paramétrico.

A comparação de modelos pode ser feita pelo Critério de Informação de Akaike (AIC), razão de verossimilhança e $R^2$, assim como obter evidências da invariância dos parâmetros dos itens usando correlações, raiz do erro quadrático médio (RMSE) e estatísticas *person-fit* para detectar padrões de respostas inconsistentes (DE AYALA, 2009).

O pressuposto da unidimensionalidade do teste supõe que existe um traço latente dominante responsável pela realização do conjunto de itens (ANDRADE; TAVARES; VALLE, 2000). Violar este pressuposto pode levar a desajustes na estimação dos parâmetros ou erros padrão elevados (DEMARS, 2010).

A especificação cuidadosa dos subdomínios do construto garante que o BI inclua todos os aspectos relevantes para mensurá-lo. No entanto, é importante verificar se alguns domínios podem ser vistos como parte de um construto global (dimensão) ou se devem ser tratados como construtos separados (dimensões) (BJORNER et al., 2007).

A avaliação da dimensionalidade faz parte da avaliação da validade do construto (THISSEN et al., 2007), a qual pode ser testada de várias formas: por análise fatorial e modelagem de equações estruturais (DE AYALA, 2009); análise fatorial com matriz de correlações tetracóricas, no caso de modelos dicotômicos (ANDRADE; TAVARES; VALLE, 2000); análise fatorial confirmatória e exploratória, teste *scree*, análise paralela (REEVE et al., 2007) e análise fatorial de informação completa (*full-information*) (BOCK; GIBBONS; MURAKI, 1988).

No entanto, para que a unidimensionalidade possa ser considerada, Reeve et al. (2007) adverte que o maior fator deve explicar pelo menos 20% da variabilidade total e a razão deste para o segundo maior fator deve ser maior do que 4. Para Lai et al. (2011), deve-se considerar unidimensionalidade quando cargas padronizadas são maiores do que 0,3 para todos os itens do fator geral.

Outro pressuposto da TRI é a independência local. Admite-se que, de acordo com o traço latente comum que está sendo medido, as respostas dos indivíduos aos itens que compõem o teste são independentes entre si (ANDRADE; TAVARES; VALLE, 2000; EDWARDS, 2009). Dependência local pode ocorrer quando um conjunto de questões se refere ao mesmo problema de pesquisa, por exemplo, questões de trigonometria que se referem à mesma figura, questões de compreensão que se referem ao mesmo trecho de leitura; quando o texto de algum item fornece informação para responder a outro item, entre outras razões (DE AYALA, 2009; MURPHY; DODD; VAUGHN, 2010).

Na prática, segundo Murphy, Dodd e Vaughn (2010), existem situações em que esta suposição de independência é improvável de se assegurar. Quando a dependência local existe, as estimativas dos parâmetros da TRI podem ser tendenciosas e

imprecisas (REEVE et al., 2007). Para testar a dependência local, a matriz de correlações residuais resultantes de uma análise fatorial confirmatória pode ser analisada; coeficientes com valores superiores a 0,2 indicam dependência (REEVE et al., 2007). De Ayala (2009) sugere utilizar o índice Q3, proposto por Yen (1984), que diz respeito a correlação entre resíduos para pares de itens.

Por fim, Thissen et al. (2007) ressaltam que não existem regras definitivas para decidir quando multidimensionalidade e dependência local possuem magnitude suficiente para causar problemas. Na prática, geralmente considera-se que, se os dados se ajustam adequadamente ao modelo da TRI, há um fator dominante responsável pelo conjunto de itens. Desta forma, cabe ao pesquisador em conjunto com profissionais especialistas de conteúdo definir se os efeitos podem ou não ser significativos para o traço latente, e se é razoável supor a unidimensionalidade.

**Estimação dos parâmetros do modelo da TRI**

A etapa de estimação dos parâmetros do modelo é muito importante, pois é neste momento que será definida a escala que servirá de referência para a interpretação de resultados do teste (BAKER, 2001). A situação mais comum nesta etapa é quando se tem apenas a resposta dos indivíduos aos itens e, a partir dessas informações, estimam-se os parâmetros dos itens e o traço latente dos respondentes, simultaneamente. Neste caso, a estimação pode ser conjunta ou em duas etapas, ou seja, primeiro a estimação dos parâmetros dos itens e, posteriormente, os traços latentes (ANDRADE; TAVARES; VALLE, 2000).

Outras situações que podem ocorrer são: quando os parâmetros dos itens são conhecidos e deseja-se estimar o traço latente; ou quando o traço latente dos respondentes são conhecidos e deseja-se estimar os parâmetros dos itens (ANDRADE; TAVARES; VALLE, 2000). O processo de estimação dos parâmetros dos itens é conhecido como calibração.

De acordo com Bjorner et al. (2007) e Thissen et al. (2007), na fase de calibração do BI e definição da escala, a representatividade da amostra (por ex., em termos de estruturas sócio-demográficas semelhantes como a população alvo) é menos

importante do que ter respostas suficientes em todos os níveis do traço latente e certa frequência de resposta em cada categoria. Portanto, grandes amostras de respondentes são necessárias para estimar com precisão os parâmetros dos itens (BJORNER et al., 2007; DEMARS, 2010). Por outro lado, em testes de alto impacto não é interessante expor desnecessariamente os itens. Mais detalhes sobre o assunto são apresentados na seção 3.1.1.

Há vários métodos de estimação, os quais podem ser classificados como clássicos (frequentistas) ou bayesianos. Para obter soluções para estes métodos, é necessário o uso de algum programa computacional. Após a definição do método e estimadores, estes devem ser mantidos ao longo do tempo.

Segundo Hambleton, Swaminathan e Rogers (1991), os métodos de estimação mais utilizados são:

- **Procedimento de máxima verossimilhança conjunta** (LORD, 1974, 1980) - os parâmetros dos itens e os traços latentes são estimados simultaneamente, exigindo um grande esforço computacional. Neste caso, é preciso estabelecer uma métrica para os parâmetros, a fim de eliminar o problema de falta de identificabilidade do modelo (ver ANDRADE; TAVARES; VALLE, 2000).
- **Procedimento de máxima verossimilhança marginal** (MML - BOCK; LIEBERMAN, 1970; BOCK; AITKIN, 1981) – estimação em duas etapas. Considera-se certa distribuição para os traços latentes e os parâmetros dos itens são estimados. Uma vez determinados os parâmetros dos itens, os traços latentes são estimados.
- **Procedimentos de estimação Bayesiana marginal e conjunta** (MISLEVY, 1986; SWAMINATHAN; GIFFORD, 1986) - distribuições *a priori* são estabelecidas para os parâmetros dos itens e traços latentes, levando em consideração suas limitações e eliminando alguns problemas como estimação indevida de parâmetros e a não-convergência. Constroe-se uma nova função denominada distribuição *a posteriori* e estimam-se os parâmetros de interesse com base em alguma

característica dessa distribuição, como a média ou a moda (ANDRADE; TAVARES; VALLE, 2000).

Para os métodos de máxima verossimilhança, aplica-se algum processo iterativo para determinar as estimativas dos parâmetros, como o algoritmo *Newton-Raphson*, algoritmo EM ou *Scoring* de Fisher. Geralmente, adotam-se as distribuições Log-normal ou Qhi-Quadrado para $a_i$; distribuição Normal para $b_i$; e distribuição Beta para $c_i$ (ANDRADE; TAVARES; VALLE, 2000).

**Estimadores**

Os métodos bayesianos de estimação incorporam informação *a priori* para os dados, enquanto a estimação pela máxima verossimilhança (MV) depende dos dados por si só (WANG; VISPOEL, 1998). Na abordagem frequentista, MV visa maximizar a função de verossimilhança; já o estimador ponderado pela verossimilhança (WLE), visa corrigir o viés do método MV, maximizando a função de verossimilhança, ponderada por uma determinada função (WARM, 1989).

Na abordagem Bayesiana, um estimador pontual do traço latente pode ser baseado na sua distribuição *a posteriori*, tais como: máxima *a posteriori* (MAP) ou Bayes modal, que utiliza a moda da distribuição *a posteriori* como estimativa do traço latente, e esperança ou média da distribuição *a posteriori* (EAP).

Esses métodos também podem ser utilizados nos testes adaptativos. Nos casos em que a utilização de MV é inviável, por exemplo, quando todas as respostas são corretas ou todas são incorretas, o método EAP pode ser utilizado para estimar o traço latente (VAN DER LINDEN; PASHLEY, 2010; HAN; GUO, 2011).

**Função de Informação do Item e do BI**

É fundamental saber quais itens são mais úteis para medir níveis particulares do traço latente quando se pretende desenvolver testes curtos ou selecionar itens em CAT. A função de informação do item (FII) em conjunto com a CCI permite analisar quanto um

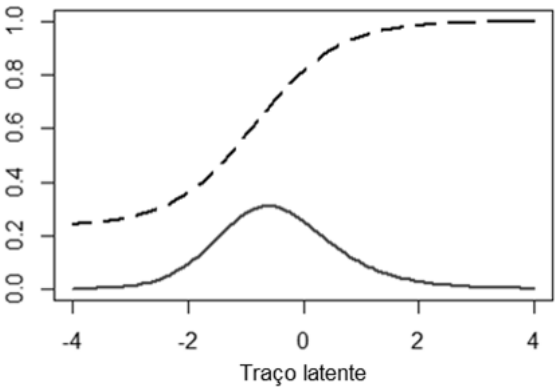
item contém de informação para a medida do traço latente (ANDRADE; TAVARES; VALLE, 2000), ou seja, a quantidade de precisão adicional que é obtida através da apresentação daquele item para qualquer nível em particular (THISSEN et al., 2007).

Um item mede o traço latente com maior precisão no nível correspondente ao parâmetro de dificuldade do item e esta quantidade de informação do item diminui à medida que o traço latente se afasta da dificuldade do item, aproximando-se de zero nos extremos da escala (BAKER, 2001).

A Figura 4 representa a CCI de um item (linha tracejada) e sua respectiva FII (linha contínua). Quanto mais alto é o pico da FII, maior é a contribuição do item para o teste naquele nível do traço latente. A altura das curvas de informação é uma função do poder de discriminação do item e a localização das curvas é determinada pelo parâmetro de dificuldade (REEVE et al., 2007).

Na Figura 4, observa-se que o item possui um máximo próximo ao nível -0,8. Assim, para métodos que fazem uso dessa informação no CAT, este item provavelmente seria selecionado para aplicação a indivíduos cuja estimação provisória do traço latente estaria próxima a este valor em algum estágio do teste.

Figura 4 – FII e CCI.



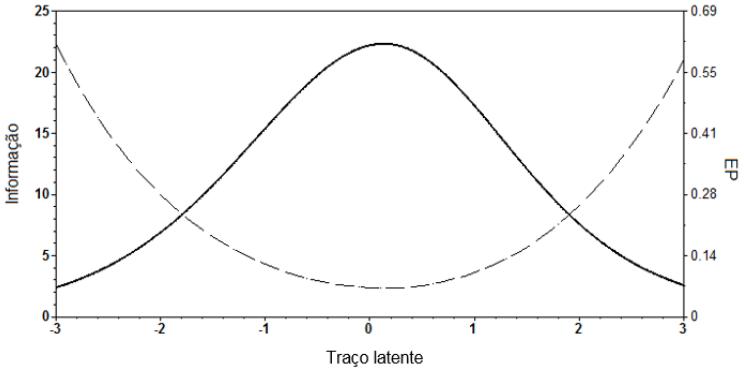

Fonte: Elaborada pela autora.

Para avaliar a totalidade da informação obtida a partir de uma combinação de itens, as FIIs são simplesmente somadas para obter a função de informação de teste (FIT), a partir da qual é possível definir a precisão do teste para um determinado nível do traço latente (ANDRADE; TAVARES; VALLE, 2000; BAKER, 2001; BJORNER et al., 2007). Logo, essas medidas são dependentes do valor do traço latente.

A Figura 5 representa a FIT de um determinado teste (linha contínua) e o erro padrão (EP - linha tracejada) obtido em toda a gama de valores da escala. Uma boa FIT deve ser mais abrangente possível, principalmente para que possa medir com precisão indivíduos localizados nos extremos da escala. Maior precisão pode ser obtida para respondentes próximos do pico da FIT.

Na escala N(0,1), o zero refere-se ao nível médio do traço latente, valores negativos ou positivos referem-se aos níveis abaixo ou acima da média. Em CAT, a FIT passa a ser denominada de FIBI - função de informação do bando de itens - porque não há mais um "teste", mas sim, há tantos testes (potenciais) quanto o número de respondentes.

Figura 5 – Função de informação do teste e erro padrão das estimativas do traço latente nos diferentes níveis da escala.




Fonte: Elaborada pela autora.

A visualização da FIBI é muito importante, pois fornece informações das regiões da escala que necessitam de mais itens para melhorar as estimativas. Além disso, auxilia na decisão sobre qual regra de parada deve ser adotada para os testes, principalmente quando é utilizado algum critério de precisão.

Nesses casos, se o BI não é capaz de estimar com alta precisão todos os níveis da escala, respondentes localizados nos extremos poderão nunca terminar o teste se o EP adotado para finalização do teste for muito baixo (SPENASSATO; BORNIA; TEZZA, 2015). Por isso, medidas para evitar este problema devem ser adotadas, como combinar diferentes regras de parada.

Segundo Edwards (2009), algumas questões devem ser consideradas ao elaborar uma escala, como: Quem estamos tentando medir? Quão confiável o traço latente precisa ser? Quão pequena é a mudança que estamos interessados em detectar? Qual é o objetivo principal desses traços latentes? O autor afirma que as respostas a estas perguntas podem ser convertidas em uma função de informação alvo que pode servir como um modelo para o processo de construção da escala e medida.

#### 2.2.2.2 *Softwares* para análises TRI

Vários *softwares* foram desenvolvidos ao longo dos anos para auxiliar na tarefa de análises dos itens para construção do BI. Zhao e Hambleton (2009) apresentam uma lista de *softwares* que já foram desenvolvidos para análises da TRI.

O Quadro 1 apresenta alguns dos *softwares* que estão disponíveis, os modelos (tipos de dados) implementados para calibração dos itens, o site onde eles estão disponíveis e se o *software* possui uma função para análise do DIF (marcados com "X"). Nos artigos analisados nesta tese, observou-se a prevalência do uso dos *softwares* Mplus, Parscale e Multilog.

Quadro 1 – *Softwares* para análises TRI.

| ***Software*** | **Modelos para estimação** | **Site** | **Licença** | **Análise DIF** |
|---|---|---|---|---|
| BILOG-MG | modelos dicotômicos | http://www.ssicentral.com/irt/ | comercial | x |
| ConQuest | modelos unidimensional e multidimensional - dicotômicos e politômicos | https://shop.acer.edu.au/group/CON3 | comercial | x |
| ConstructMap | modelos unidimensional e multidimensional | http://bearcenter.berkeley.edu/software/constructmap | livre | x |
| EQSIRT | modelos unidimensional e multidimensional - dicotômicos e politômicos | http://www.mvsoft.com/eqsirt10.htm | comercial | x |
| FACETS | modelos multi-facetas | http://www.winsteps.com/facets.htm | comercial (tem versão demo) | |
| flexMIRT | modelos unidimensional e multidimensional - dicotômicos e politômicos | https://www.vpgcentral.com/irt-software/ | comercial | x |
| IRTPRO | modelos dicotômicos e politômicos | http://www.ssicentral.com/irt/ | comercial | x |
| MPLUS | modelos dicotômicos e politômicos | https://www.statmodel.com/index.shtml | comercial | x |
| MULTILOG | modelos politômicos | http://www.ssicentral.com/irt/ | comercial | |

Continuação

| *Software* | Modelos para estimação | Site | Licença | Análise DIF |
|---|---|---|---|---|
| NOHARM | modelos unidimensional e multidimensional -dicotômicos | http://noharm.software.informer.com/ | livre | |
| PARAM | modelos dicotômicos | http://echo.edres.org:8080/irt/param/ | livre | |
| PARSCALE | modelos dicotômicos e politômicos | http://www.ssicentral.com/irt/ | comercial | x |
| RUMM | modelo de Rasch | http://www.rummlab.com/ | comercial | x |
| SAS | vários macros para TRI: IRT-FIT, PROC NLMIXED, IRT-Lab, GLIMMIX, IRTGEN | http://www.sas.com/pt_br/home.html | comercial | x |
| Stata | modelos dicotômicos e politômicos | http://www.stata.com/stata14/irt/ | comercial | x |
| TESTfact | análise clássica e fatorial | http://www.ssicentral.com/ | comercial | |
| WinGen | geração de dados - parâmetros e respondentes para modelos unidimensionais e multidimensionais | http://www.umass.edu/remp/software/simcata/wingen/homeF.html | livre | x |
| Winsteps | modelos dicotômicos e politômicos | http://www.winsteps.com/index.htm | comercial (tem versão demo) | x |
| Xcalibre | modelos dicotômicos e politômicos | http://www.assess.com/xcart/ | comercial | x |

Fonte: Elaborado pela autora

O *software* livre R (R CORE TEAM, 2016) possui várias bibliotecas que podem ser utilizadas para análises da TRI. No site do R é possível encontrar uma lista de pacotes referentes a modelos e métodos psicométricos. Alguns pacotes que podem ser utilizados são:

- **Estimação de parâmetros da TRI** - eRm (MAIR; HATZINGER; MAIER, 2015), mRm (PREINERSTORFER, 2013), ltm (RIZOPOULOS, 2006), TAM (KIEFER; ROBITZSCH; WU, 2015), mirt (CHALMERS, 2012), mcIRT (REIF, 2014), sirt (ROBITZSCH, 2015), pcIRT (HOHENSINN, 2015), psychomix (FRICK et al., 2012) e irtoys (PARTCHEV, 2014);
- **Análise do ajuste do modelo, pressupostos e *plot*:** mokken (VAN DER ARK, 2012), KernSmoothIRT (MAZZA; PUNZO; MCGUIRE, 2014), psych (REVELLE, 2015), mirt e sirt;
- **Análise do DIF**: mirt, sirt, difR (MAGIS; BELAND; RAICHE, 2015), lordif (CHOI; GIBBONS; CRANE, 2014) e DFIT (CERVANTES, 2014);
    - **Equalização:** equateIRT (BATTAUZ, 2014), kequate (ANDERSSON; BRANBERG; WIBERG, 2013), SNSequate (GONZALEZ, 2014) e irtoys.

### 2.2.3 Implantação de um CAT

Testes adaptativos computadorizados são compostos, basicamente, por cinco componentes: **(1)** conjunto de itens calibrados pela TRI (banco de itens), **(2)** método para iniciar o teste, **(3)** método de seleção dos itens, **(4)** método de estimação do traço latente e **(5)** regra para finalização do teste. Diversas restrições podem ser impostas junto ao método de seleção de itens.

De acordo com Thompson e Weiss (2011), muitas pesquisas têm sido realizadas ao longo dos últimos 40 anos sobre os aspectos técnicos de um CAT, como algoritmos de seleção de itens, controles de exposição de itens e regras de parada do teste. Desta forma, há uma extensa literatura sobre comparação de métodos

para cada componente de um CAT, cujo objetivo é obter os melhores resultados para um teste em particular.

A especificação de restrições é imposta principalmente em testes de alto impacto, pois há um maior controle dos itens para não comprometer os resultados. No entanto, é preciso tomar cuidado com a quantidade de restrições em um teste, pois conforme Luecht, de Champlain, Nungester (1998) e van der Linden (1999), restrições desnecessárias são caras e devem ser evitadas porque podem levar à redução da precisão e eficiência do teste, principalmente, quando há maior número de restrições do que há itens. Além disso, van der Linden (1999) destaca que a presença de muitas restrições retarda a convergência inicial do estimador do traço latente.

Uma implementação bem sucedida do CAT depende da acurácia dos métodos estatísticos usados para estimar o traço latente e da eficiência do método de seleção de itens (CHENG; LIOU, 2000), além de um BI com boas características psicométricas. A escolha do método mais adequado para cada componente do CAT depende do contexto do teste. As decisões tomadas nesta fase inicial de implementação do CAT são muito importantes para o resultado do teste na prática, por isso, métodos são comparados por meio de simulações.

#### 2.2.3.1 Método para iniciar o CAT

Nesta primeira etapa é preciso estabelecer uma regra para iniciar o teste, fornecendo uma estimativa provisória inicial do traço latente do respondente. Quanto mais acurada é a estimativa inicial, mais apropriada será a seleção dos próximos itens (CHEN; ANKENMANN; CHANG, 2000).

Geralmente, adota-se um nível inicial mediano, que pode ser fixo, centrado na média ou um valor aleatório dentro de um intervalo mediano para não impactar negativamente nas estimativas do traço latente (VELDKAMP; MATTEUCCI, 2013; WISE; KINGSBURY, 2000; MILLS; STOCKING, 1995). Ou ainda, com base em informações conhecidas *a priori* sobre o respondente, como o traço latente de um teste feito anteriormente,

ou outros tipos de variáveis que podem estar relacionadas com a característica medida (VAN DER LINDEN, 1999; WISE; KINGSBURY, 2000).

Weissman (2006) e Chen, Ankenmann e Chang (2000) reiteram que a estimação provisória do traço latente é tipicamente imprecisa (possui grande EP), tendenciosa (viesada) ou ambas, nos estágios iniciais de uma aplicação CAT. A seleção de itens, por sua vez, é dependente dessa estimativa, podendo apresentar itens que são incompatíveis com o verdadeiro traço latente do respondente (WEISSMAN, 2006).

Quando os respondentes começam com a mesma estimativa inicial do traço latente, os itens perto deste valor têm maior chance de se tornarem superexpostos; se a estimativa inicial está mais perto possível de seu valor verdadeiro, a exposição de itens torna-se mais distribuída ao longo da escala (VAN DER LINDEN, 1999). Uma alternativa para obter maior precisão inicial é aplicar alguns itens iniciais e posteriormente estimar o traço latente (ver VAN KRIMPEN-STOOP; MEIJER, 1999, 2001).

O Quadro 2 apresenta alguns métodos frequentemente utilizados para iniciar o CAT. Alguns deles iniciam com a combinação de métodos, por exemplo, seleção aleatória dentro de um intervalo predefinido para a dificuldade dos itens.

Aplicações na área da saúde têm utilizado o mesmo item inicial para todos os respondentes, o qual apresenta uma dificuldade mediana e, na maioria das vezes, define o construto. Luecht, De Champlain e Nungester (1998) utilizaram o método de seleção aleatória do primeiro item, justificando que muito pouco se sabe sobre o traço latente do respondente no início do teste, por isso sugerem iniciar o CAT sem informação alguma.

O estudo de Chen, Ankenmann e Chang (2000) sugere modificar a função de informação para que seja considerada a incerteza sobre as estimativas do traço latente nos níveis iniciais e, posteriormente, quando a estimativa já está próxima do valor verdadeiro, pode-se utilizar o método MFI.

Chang e Ying (1999) sugerem o uso de uma estratégia de seleção de itens que estratifica o BI e usa itens menos discriminativos no início do teste, para que os itens com elevada

discriminação sejam usados em fases posteriores para obter maior precisão na estimativa final do traço latente.

Quadro 2 – Exemplos de regras para iniciar o CAT.

| Método | Exemplos | Referências |
|---|---|---|
| Seleção aleatória do(s) primeiro(s) item(s) | Três itens fáceis selecionados aleatoriamente; dois itens quaisquer selecionados aleatoriamente; seleção aleatória de um item entre os cinco mais informativos para determinado nível. | Luecht, de Champlain e Nungester (1998); Wang e Vispoel (1998); Zitny et al. (2012) |
| Mesmo item para todos os respondentes | Todos respondem ao mesmo item, geralmente de dificuldade mediana. | Cella et al. (2007); Fliege et al. (2009); Smits, Cuijpers e van Straten (2011); Anatchkova et al. (2012) |
| Seleção baseada no traço latente estabelecido ou nível de dificuldade | Dois itens mais informativos para $\theta = 0$; um item com base em informações *a priori* sobre o traço latente do respondente; um item com $b = 0$ ou vários itens com diferentes níveis de dificuldade. | Chang e Ying (1996); Vispoel (1998); van der Linden, Scrams e Schnipke (1999); Meijer (2002); Weissman (2006); Ozyurt et al. (2012) |

Fonte: Elaborado pela autora.

#### 2.2.3.2 Método de seleção dos itens

O método de seleção de itens adotado em CAT determina quais itens serão aplicados a cada respondente. As estatísticas do item calibrado juntamente com o conteúdo e outras propriedades qualitativas servem como principais entradas fixas para o

mecanismo de seleção usado na maioria dos *softwares* para CAT (LUECHT; DE CHAMPLAIN; NUNGESTER, 1998).

Métodos de seleção de itens têm recebido atenção considerável dos estudiosos da área. De acordo com Veldkamp e Matteucci (2013), ao longo dos últimos dez anos, vários estudos comparativos foram realizados na tentativa de encontrar o melhor método de seleção de itens, porém, não se encontrou vencedores, uma vez que a maioria deles executa muito bem quando 20 itens ou mais são selecionados para o teste.

O procedimento de seleção de itens é uma componente chave do CAT, que pode aumentar a qualidade e a eficiência do teste aplicando itens ótimos sequencialmente para cada respondente (CHOI; SWARTZ, 2009). Idealmente, os procedimentos de seleção de itens devem equilibrar duas exigências concorrentes: preservar a precisão nas estimativas do traço latente e utilizar eficazmente todos os itens de um BI (CHENG; LIOU, 2003).

A seguir, apresentam-se alguns métodos de seleção de itens. Esses métodos estão atualmente implementados no pacote catR (MAGIS; RAICHE, 2012) do *software* R (R CORE TEAM, 2016).

**1)** ***Maximum Fisher information*** (MFI - máxima informação de Fisher) - Este método busca maximizar as informações obtidas sobre o respondente a fim de minimizar o erro da estimativa. Desta forma, seleciona o próximo item como sendo aquele que maximiza a função de informação do item (MAGIS; RAICHE, 2012).

**2) Método *bOpt*** (URRY, 1970) - consiste em selecionar o próximo item cujo nível de dificuldade é o mais próximo da estimativa atual do traço latente. Sob o modelo de um parâmetro da TRI, os métodos bOpt e MFI são equivalentes.

**3) Método *thOpt*** (BARRADA; MAZUELA; OLEA, 2006) - seleciona os itens cujo valor de $\theta$ ótimo (isto é, para o traço latente onde a informação de Fisher é máxima) está tão perto quanto possível da estimativa atual do traço latente. Para os modelos de um e dois parâmetros da TRI, os métodos bOpt e thOpt são equivalentes.

**4)** ***Maximum likelihood weighted information*** (MLWI) (VEERKAMP; BERGER, 1997) - seleciona o próximo item como

sendo o único com máxima informação ponderada pela função de verossimilhança.

**5)** ***Maximum posterior weighted information*** (MPWI) (VAN DER LINDEN, 1998) - seleciona o próximo item como sendo o único com máxima informação ponderada pela distribuição *a posteriori*.

**6)** ***Maximum expected information*** (MEI) (VAN DER LINDEN, 1998) - seleciona o item com a máxima informação esperada.

**7)** ***Minimum expected posterior variance*** (MEPV) (CHOI; SWARTZ, 2009; OWEN, 1975; VAN DER LINDEN, 1998) - seleciona o item com a mínima variância posterior esperada.

**8)** ***Kullback-Leibler*** (KL) (CHANG; YING, 1996; BARRADA et al., 2010) - seleciona o item com máxima informação KL em torno do traço latente estimado, ou seja, fornece uma informação global. A função de informação KL ponderada pela verossimilhança avalia a capacidade de discriminação do item entre quaisquer possíveis pares de traços.

**9)** ***Posterior Kullback-Leibler*** (KLP) (CHANG; YING, 1996; BARRADA et al., 2010) - seleciona o item com máxima informação KL ponderada pela distribuição *a posteriori* do traço latente.

**10)** ***progressive method*** (PG) (BARRADA et. al 2008, 2010; REVUELTA; PONSODA, 1998) - o item selecionado é aquele que maximiza a soma de dois elementos, uma parte aleatória e uma parte determinada pela informação de Fisher. No início do teste, a importância do elemento aleatório é máximo; com o avanço no teste, a informação aumenta sua relevância na seleção do item. A velocidade de transição da seleção puramente aleatória para seleção completamente baseada na informação é determinada por um parâmetro de aceleração (AP), onde valores mais altos implicam uma maior importância do elemento aleatório durante o teste.

**11)** ***proportional method*** (PP) (BARRADA et al., 2008, 2010; SEGALL, 2004) - os itens são selecionados aleatoriamente com probabilidades de seleção determinadas por suas informações de Fisher elevado a uma determinada potência. Esta potência é igual

a zero no início do teste e aumenta à medida que o teste avança. Isto implica que o teste começa com a seleção completamente aleatória e se aproxima da MFI no final do teste. Neste método, AP desempenha um papel semelhante ao do método PG.

**12)** ***random*** - a seleção do próximo item é completamente aleatória entre os itens disponíveis.

Vários estudos compararam métodos de seleção de itens e seus impactos nas estimativas do traço latente dos respondentes. O Quadro 3 apresenta alguns deles. Destaca-se que os métodos foram testados em situações particulares de testes, ou seja, com diferentes composições e tamanhos do BI, regras do CAT e restrições no algoritmo, distribuições do traço latente, número de respondentes, etc. Dessa forma, é difícil afirmar qual deles é o mais eficiente para todas as situações de testes, tornando-se necessário efetuar simulações e compará-los.

#### 2.2.3.3 Restrições na seleção dos itens

A necessidade de impor restrições em CATs depende dos objetivos do teste. Essas restrições podem ser estatísticas, ou não, e muitas vezes são impostas, por exemplo, para que o teste seja adequadamente aplicado, para evitar que itens com pistas para outros itens estejam no mesmo teste e para medir com eficiência todos os domínios do traço latente que está sendo investigado (VAN DER LINDEN, 1998; 2010; ZHENG, 2014). Essas restrições podem ser estabelecidos diretamente no BI ou no algoritmo de seleção de itens (VAN DER LINDEN, 2010).

O Quadro 4 sintetiza alguns métodos mais utilizados para impor restrições no algoritmo. Em van der Linden e Reese (1998) é possível encontrar uma lista de possíveis restrições em testes. As restrições mais utilizadas em testes estão relacionadas ao balanceamento de conteúdo (BC), controle da taxa de exposição do item e restrições referentes ao tempo de resposta ao item e de teste.

Quadro 3 – Estudos comparativos de métodos de seleção de itens para CAT.

| Autor | Métodos | Resultados |
|---|---|---|
| van der Linden (1998) | MPWI; MEI; MEPV; MEPWI; MFI | Sugere-se o uso de MEI, MEPV ou MEPWI por apresentarem menor viés após 10 itens e se manterem superiores aos demais métodos para testes mais longos. MFI apresentou o pior desempenho para testes curtos. |
| Chen, Ankenmann e Chang (2000) | MFI, IIF (Informação Intervalar de Fisher), MPWI, KL e KLP | Para testes com mais de 10 itens, não houve vantagem significativa de qualquer um dos métodos em relação à precisão. IIF, MPWI, KL e KLP foram marginalmente melhores do que MFI nos estágios iniciais de CAT para traços latentes igual a -3 e -2. |
| Chang e Ying (1996) | KL e MFI | KL melhora o viés e reduz o erro quadrático médio, principalmente nos estágios iniciais do teste. Sugere-se o uso de KL quando poucos itens forem aplicados e, posteriormente, MFI pode ser utilizada. |
| Barrada et al. (2010) | MFI, MLWI, KL, PG, PP e máxima informação estratificada com blocos. | Os autores compararam seis métodos de seleção de itens em relação ao RMSE e taxa de sobreposição de itens, variando a taxa máxima de exposição para testes com 20 e 40 itens. Eles concluíram que os melhores métodos foram o KL e PP. |
| Veerkamp e Berger (1997) | MFI, FII e MLWI | Os autores compararam os três métodos de seleção de itens com estimadores MV e EAP para o traço latente. Os resultados mostraram que MLWI é uma boa alternativa para MFI, assim como utilizar MFI com estimador EAP. |
| Costa et al. (2009) | MFI, KL, MEI | Os autores utilizaram o método EAP para estimar o traço latente e concluíram que os três métodos de seleção de itens apresentaram um comportamento semelhante em relação ao viés e RMSE. |

Fonte: Elaborado pela autora.

Quadro 4 – Principais métodos para inserir possíveis restrições em CAT.

| Método | Tipo de restrição | Referências | Como funciona |
|---|---|---|---|
| Programação linear 0-1 (LP) | Várias | van der Linden e Reese (1998) | Maximiza a informação das estimativas do traço latente, sujeito às restrições, para selecionar os itens; o modelo LP é atualizado a cada passo permitindo a montagem de testes ótimos, garantindo que todas as especificações sejam atendidas. |
| *Shadow test approach* (STA) | Várias | van der Linden e Reese (1998); Belov, Armstrong e Weissman (2008); van der Linden (2010); He, Diao e Hauser (2014) | Este método emprega uma abordagem de otimização sequencial restrita que trata especificações de teste como restrições que devem ser impostas na seleção do item. Um item é selecionado para aplicação através da resolução simultânea de uma sequência de problemas de otimização. O item a ser aplicado é selecionado a partir do *shadow test* montado para ser o ideal para o traço latente atual do respondente e não diretamente do BI. |
| Método Sympson-Hetter (SH) | Exposição do item | Sympson e Hetter (1985); Segall (2004, 2005); Barrada, Abad e Veldkamp (2009) | Um parâmetro de controle de exposição é atribuído para cada item. Antes de um item selecionado ser aplicado, um experimento de probabilidade aleatória é conduzida pelo cálculo de uma probabilidade condicional para determinar se o item selecionado deve ser aplicado ao respondente atual. P(A/S) é a probabilidade de aplicar um item após escolhido. Se o número gerado aleatoriamente estiver abaixo desse valor, o item selecionado é aplicado. P(A/S) é constantemente atualizado. |

Continuação

| Método | Tipo de restrição | Referências | Como funciona |
|---|---|---|---|
| Estratificação do banco de itens por domínios | Conteúdo | Yi e Chang (2003); van der Linden (2010) | Estratifica o banco de itens de acordo com os domínios de conteúdo e seleciona um número de itens por domínio. |
| Multi-estágios | Exposição do item | van der Linden e Reese (1998); van der Linden (2010) | Os respondentes realizam uma sequência de subtestes, podendo avançar para um subteste mais difícil ou retornar para um mais fácil, dependendo da resposta dada. Em vez de adaptar o teste aos indivíduos, item por item, o teste se adapta aos respondentes em estágios. Cada estágio consiste em um conjunto fixo de itens e difere na dificuldade média. |
| *a*-estratificado | Exposição do item | Barrada, Mazuela e Olea (2006); Chang e Ying (1999, 2009) | Itens menos discriminativos são selecionados e aplicados no início do teste, quando a estimativa é menos precisa. Itens altamente discriminativos são aplicados em estágios mais avançados. |
| Elegilidade do item (IE) | Exposição do item | Barrada, Abad e Veldkamp (2009) | É uma modificação no método SH. O controle da exposição é feito por meio da restrição da proporção de respondentes para qual um item pode ser elegível. Para cada respondente, um subconjunto de itens elegíveis é formado antes de qualquer item ter sido administrado. Durante a aplicação, apenas itens deste subconjunto podem ser administrados. |

Fonte: Elaborado pela autora.

Estudos de Belov, Armstrong e Weissman (2008), van der Linden e Reese (1998), van der Linden (2005) e He, Diao e Hauser (2014) compararam diferentes métodos para inserir restrições mais severas em CAT. De uma forma geral, as abordagens *Shadow test* (STA) e programação linear 0-1 (LP) mostraram bons resultados em relação à exposição dos itens e à precisão, se comparadas à outros métodos; por exemplo, testes baseados em multi-estágios ou *testlet*, *weighted deviation model*, *weighted penalty model*, *maximum priority index*.

Os pesquisadores van der Linden e Reese (1998) analisaram possíveis efeitos de um grande número de restrições (até 433) nas estimativas dos traços latentes e concluíram que os efeitos são mínimos após a aplicação de 20 itens ou mais. Assim, grandes BI para atender as restrições do modelo e testes mais longos são necessários para minimizá-los (VAN DER LINDEN; SCRAMS; SCHNIPKE, 1999).

A seguir, destacam-se brevemente as principais restrições citadas neste trabalho e que são de suma importância para CATs de alto impacto. Os estudos têm por objetivo mostrar o papel que elas exercem sobre os resultados do teste, além de sua relação com a interpretação do traço latente.

#### *2.2.3.3.1 Balanceamento de conteúdo*

A restrição de balanceamento de conteúdo (BC) garante que itens de todos os conteúdos ou domínios sejam aplicados aos respondentes. Esta estratégia controla o número de itens selecionado de cada domínio, além de maximizar a informação dentro dos domínios para selecionar o próximo item a ser aplicado (LUECHT; DE CHAMPLAIN; NUNGESTER,1998).

Ao contrário de testes P&P, onde os itens são fixos e devem medir todos os domínios do traço latente, em CAT, se a restrição de conteúdo não for implementada no algoritmo, não haverá garantias de que itens de todos os domínios serão aplicados a todos os respondentes, principalmente se o método de seleção de itens busca maximizar a informação para o traço latente estimado (VAN DER LINDEN, 2010).

Foram realizados alguns estudos a fim de demonstrar a importância do BC e seus impactos na avaliação do traço latente quando esta restrição não é imposta. Em Fliege et al. (2009), os resultados indicaram que, sem BC, apenas 42% do total dos itens disponíveis foram utilizados, impactando na representatividade dos domínios que deveriam ser avaliados e no uso dos itens disponíveis.

Zheng, Chang e Chang (2013) compararam a seleção aleatória de itens e a seleção por MFI, com e sem BC. Os resultados mostraram que MFI com BC estimou melhor o traço latente; além disso, itens de um domínio com maior poder de discriminação foram mais selecionados, apresentando altas taxas de exposição.

Luecht, de Champlain e Nungester (1998) avaliaram os impactos considerando: (1) um domínio mais difícil do que outro (dois domínios) e (2) alguns domínios com menos informação do que outros (cinco domínios), além de diferentes níveis de dificuldade. Os resultados mostraram que o traço latente e a precisão não foram afetados significativamente, mas a validade de conteúdo pode ser comprometida quando o BC é ignorado, gerando superexposição de itens que pertencem às áreas mais informativas e o gasto desnecessário com o desenvolvimento de itens que não são utilizados.

Especialmente em CATs de alto impacto, a superexposição de itens é uma grande preocupação porque quanto maior ela é, maior é o risco à segurança do CAT, podendo ameaçar a validade do teste caso exista uma vantagem adquirida por um respondente devido ao conhecimento prévio dos itens (REVUELTA; PONSODA,1998; BARRADA; ABAD; VELDKAMP, 2009; LEROUX et al., 2013).

#### *2.2.3.3.2 Taxa de exposição dos itens e de sobreposição de teste*

Uma abordagem para reduzir o risco à segurança do teste é por meio do controle da taxa de exposição dos itens (FINKELMAN NERING; ROUSSOS, 2009) e de sobreposição de testes.

A taxa de exposição é um índice que se refere a proporção de vezes que um item é aplicado sobre o total de testes CATs que são aplicados; já a taxa de sobreposição de teste é definida como a proporção de itens que, em média, são compartilhados por todos os pares possíveis de respondentes (CHEN et al., 2014; CHEN, 2010; MILLS; STOCKING, 1995; CHEN; ANKENMANN; SPRAY, 2003, CHEN; LEI, 2005, WAY, 1998).

Vários métodos têm sido propostos para reduzir essas taxas, bem como aumentar o uso dos itens do BI e melhorar sua segurança, buscando manter a precisão do teste (VEERKAMP; GLAS, 2000; BARRADA; ABAD; VELDKAMP, 2009; LEROUX et al., 2013; VELDKAMP; MATTEUCCI, 2013).

Conforme Georgiadou, Triantafillou e Economides (2007), os métodos de controle de exposição podem ser classificados em cinco grupos de estratégias: procedimentos de seleção condicionais, aleatórios, estratificados, métodos combinados e *design* de testes adaptativos em multi-estágios (Quadro 5). Os autores apresentam uma revisão de métodos de 1983 a 2005.

Métodos sofisticados de controle estatístico das taxas de exposição de itens surgiram a partir dos esforços de Sympson e Hetter (1985), após a segunda metade da década de 1990 (STOCKING; LEWIS, 2000). Esses métodos incluem vários tipos de condicionamento no controle estatístico, como mostram os trabalhos de Davey e Parshall (1995), Nering, Davey e Thompson (1998), Stocking e Lewis (1998), van der Linden (2003), entre outros.

Segundo Stocking e Lewis (2000), procedimentos condicionais ao traço latente, como de Stocking e Lewis (1998), permitem o controle direto da exposição do item para diferentes níveis do traço latente, podendo ser possível escolher diferentes taxas máximas de exposição alvo para diferentes níveis do traço latente.

Por outro lado, os procedimentos condicionais aos itens, como de Davey e Parshall (1995), procuram controlar as probabilidades de exposição dos itens condicionais aos itens que já apareceram no teste. Isso pode ser visto como um método de tentar

controlar a sobreposição de teste, além de limitar a frequência de uso do item (STOCKING; LEWIS, 2000).

Quadro 5 – Métodos de controle da taxa de exposição dos itens.

| Método | Objetivo | Exemplo |
|---|---|---|
| **Procedimentos condicionais** | Visam controlar as taxas de exposição com base em algum critério. | Frequência de uso – método SH (SYMPSON; HETTER, 1985); Elegibilidade do item - IE (VAN DER LINDEN; VELDKAMP, 2007), condicionais aos níveis do traço latente (STOCKING; LEWIS, 1998) |
| **Procedimentos aleatórios** | Introduzem alguma randomização no processo de seleção de itens para um determinado subconjunto de itens. | Métodos 5-4-3-2-1 (MCBRIDE; MARTIN, 1983) e *randomesque* (KINGSBURY; ZARA, 1989) |
| **Métodos estratificados** | Buscam estratificar o conjunto de itens de acordo com propriedades estatísticas e os itens são aplicados a partir de um determinado estrato. | Métodos *a*-estratificado (CHANG; YING, 1999); *a*-estratificado com *b*-blocos (CHANG; QIAN; YING, 2001); *a*-estratificado com blocos de conteúdos (YI; CHANG, 2003); *maximum information stratification with blocking* (BARRADA, MAZUELA, OLEA, 2006) |
| **Procedimentos combinados** | Quando dois ou mais métodos de controle de exposição são combinados. | Método progressivo restrito (REVUELTA; PONSODA, 1998); Erro padrão progressivo restrito (MCCLARTY; SPERLING; DODD, 2006). |

Fonte: Adaptado de Leroux et al. (2013).

Vários estudos têm comparado métodos de controle da exposição de itens em busca do melhor método (ver DAVEY; PARSHALL, 1995; REVUELTA; PONSODA, 1998; BARRADA

et al., 2008; CHANG; ANSLEY, 2003; BARRADA; OLEA; PONSODA, 2007; VAN DER LINDEN, 2003; BARRADA; ABAD; VELDKAMP, 2009). Porém, Revuelta e Ponsoda (1998), Georgiadou, Triantafillou e Economides (2007) e Leroux et al. (2013) destacam que cada um deles tem suas particularidades, vantagens e desvantagens, não podendo ser generalizado para todas as situações.

**Controle de ambas taxas: exposição e sobreposição**

A maioria dos estudos comparam apenas os métodos de controle da exposição dos itens e verificam seus impactos na taxa de sobreposição de teste. No entanto, quando a exposição do item for controlada incondicionalmente, a taxa de exposição global de um item pode ser baixa para respondentes em toda a escala de medida, mas este item pode ter sido administrado a quase todos os respondentes em um nível específico (BARRADA et al., 2009; CHEN; LEI, 2005).

Nesse sentido, a exposição do item e a sobreposição de testes em traços semelhantes é uma preocupação legítima para qualquer programa CAT, pois favorecem o pré-conhecimento de itens (WAY, 1998), principalmente porque respondentes podem obter informações sobre o teste de mais de um respondente (fonte), conseguindo maior alcance de partilha de informação (CHEN; LEI, 2010).

Chen, Ankenmann e Spray (2003) e Chen e Lei (2005) reiteram que métodos condicionais apenas a nível do item, por mais que reduzam a sobreposição de teste, não conseguem controlar tal taxa, sendo necessário utilizar métodos que visam controlar ambas as taxas. Alguns estudos vêm sendo desenvolvidos nessa direção, como Chen e Lei (2005), Chen, Lei e Liao (2008), Chen e Lei (2010), Chen (2010) e Chen et al. (2014).

Chen e Lei (2005) propuseram um método para CAT de comprimento fixo, que é a extensão do método SH, denominado SHT (*SH Procedure with Test Overlap Control*). Para controlar ambas as taxas simultaneamente, o procedimento tenta controlar não só o valor máximo, mas também a variância das taxas de exposição do item. Assim, simulações iterativas para definir esses

parâmetros de exposição são repetidos até se obter valores inferiores aos preestabelecidos. Para assegurar a representação de conteúdo adequada em CATs, este procedimento pode ser implementado dentro de cada área de conteúdo em vez de todo o BI.

Os autores Chen e Lei (2005) compararam os métodos SH, SLC (método multinomial condicional de Stocking e Lewis (1998)).e SHT em relação ao controle de exposição do item e na precisão das estimativas. Considerando uma taxa de sobreposição máxima rigorosa, o procedimento SHT desempenhou melhor do que SLC no controle da exposição. Porém, com uma taxa menos rigorosa, o procedimento SHT desempenhou de forma semelhante a SH. De forma geral, os autores concluíram que essas taxas podem ser controladas simultaneamente em SHT, e que o controle de exposição do item melhorou, bem como a precisão das estimativas do traço latente.

Por outro lado, Chen, Lei e Liao (2008) destacam que o método SHT precisa da definição dos parâmetros de exposição, o que não é muito trivial, e esses parâmetros precisam ser reconduzidos sempre que houverem mudanças nas configurações do CAT (por ex. BI, método de seleção de item, exclusão de item, etc.) ou populações de interesse.

Para melhorar este problema, Chen, Lei e Liao (2008) propuseram uma versão on-line do método SHT, denominada SHTO, onde parâmetros de exposição são atualizados sequencialmente em tempo real. Esse método visa controlar as taxas sem utilizar simulações iterativas que são demoradas. Os autores comparam SHTO com duas versões on-line alternativas e SHT. Os resultados indicaram que o método SHTO é capaz de controlar eficientemente as taxas de exposição do item e de sobreposição de teste, proporcionando melhor controle de segurança geral. Porém, o método não garante o controle para um determinado nível do traço latente, sugerindo uma extensão deste, com o método condicional SLC.

Conforme Chen (2010), o tipo de sobreposição de teste considerado nos procedimentos SHT e SHTO é a sobreposição *pairwise* de teste, definido como a proporção de itens comuns

compartilhados por pares de respondentes. Porém, na prática os respondentes podem obter informações de teste de mais de um respondente anterior, e esta informação de maior compartilhamento precisa ser levada em conta nos procedimentos de controle de exposição, sendo necessário controlar a proporção de itens comuns que podem ser compartilhados por um grupo de respondentes.

Nesse contexto, Chen (2010) propõe um novo método de controle de sobreposição de teste, o procedimento SH com controle de sobreposição de teste geral (SHGT), que é capaz de controlar a proporção de itens comuns entre um respondente e um grupo anterior de respondentes. O procedimento SHGT é projetado para controlar a taxa de sobreposição de teste geral e taxas de exposição do item simultaneamente, sem qualquer tipo simulações iterativas realizadas antes dos CATs operacionais.

No estudo, o método foi comparado com outros dois procedimentos on-line em relação ao controle da sobreposição de teste e precisão da medição. Os resultados mostraram que o procedimento proposto é um método eficiente para controlar tanto a exposição do item quanto a sobreposição geral do teste. No entanto, há uma diminuição da precisão (CHEN, 2010).

Com o exposto, evidencia-se que os métodos são promissores no controle da exposição e sobreposição de testes, oferecendo maior segurança ao BI, ao preço de uma redução na precisão. Estas são importantes restrições para testes de alto impacto, mas a viabilidade da implantação do controle de ambas taxas precisa ser investigada pelas organizações de testes.

#### *2.2.3.3.3 Definição da taxa máxima de exposição dos itens e tamanho do BI*

Fatores determinantes da exposição de um item são o modelo da TRI utilizado, o método de seleção de itens e como o controle é inserido no algoritmo (WAY, 1998), assim como das suas propriedades psicométricas, dos outros itens que estão disponíveis no BI e da distribuição do traço latente dos respondentes (REVUELTA; PONSODA, 1998).

O controle sobre a variação das taxas de exposição de itens pode ser alcançado através da fixação da taxa máxima de exposição ($r^{max}$) (CHEN; ANKENMANN; SPRAY, 2003). Não há uma regra fixa para sua definição.

Conforme Stocking (1994), a definição da taxa é influenciada pela quantidade de respondentes que se submetem ao teste e quantas vezes o teste é aplicado ao longo do tempo. Assim, um item pode estar comprometido quando muitos respondentes veem os itens no mesmo dia, mesmo que o teste não seja administrado novamente dentro do mesmo ano, ou quando um grupo pequeno de respondentes participa dos testes, mas este é oferecido várias vezes no ano (MILLS; STOCKING, 1995).

Por exemplo, considere que um item é aplicado a não mais do que 10% de uma população de respondentes, mas existem 1.000.000 de respondentes em um ano; então, talvez a aplicação de um item para 100.000 respondentes irá comprometer a segurança do teste (STOCKING, 1994).

Taxas máximas mais utilizadas na literatura variam de 15% a 30% (ver ABAD et al., 2010; STOCKING, 1994; ALI, CHANG, 2014; BARRADA et al., 2009). Way (1998) propõe que o BI para testes de alto impacto deve fornecer uma taxa média de exposição dos itens variando entre 0,08 e 0,12; que a porcentagem média de sobreposição de testes fique entre 10-15% e o percentual médio de sobreposição de itens condicional ao traço latente não utrapasse 30%.

Atenção especial deve ser dada aos repetidores de testes, ou seja, indivíduos que optam por fazer o teste mais de uma vez e, se os mesmos itens aparecem em vários testes, eles podem ter a vantagem de pré-conhecimento dos itens. A memorização desses itens para divulgação é uma grande preocupação em CAT, principalmente se o BI é pequeno (VAN DER LINDEN; SCRAMS; SCHNIPKE, 1999; MCLEOD; LEWIS, 1999).

Para evitar este problema, deve-se inserir o bloqueio de itens previamente vistos para que não sejam novamente selecionados; a não ser que o CAT seja curto e o BI grande o suficiente para reduzir a exposição dos itens ao ponto de o bloqueio ser desnecessário (WAY, 1998).

Uma questão chave para reduzir o risco à segurança do teste é ter um grande BI com níveis de dificuldade bem distribuídos ao longo da escala. Segundo Davey (2011), na prática o tamanho do BI depende de uma variedade de fatores, que acaba sendo determinado pelo peso entre os benefícios de bancos de itens maiores (maior eficiência e mais segurança) contra os custos práticos e financeiros do desenvolvimento e pré-teste de um número maior de itens.

Estudos apontam vários tamanhos ideais do BI para CAT, os quais geralmente variam de 5 a 10 vezes o número de itens a serem aplicados para os respondentes (STOCKING, 1994; DAVEY, 2011; OZYURT et al., 2012). Way (1998) ressalta que, em geral, o tamanho do BI é muito influenciado pela necessidade de taxas aceitáveis de exposição de item e sobreposição de testes, que por sua vez, devem considerar os impactos associados aos resultados do teste. Contudo, o autor sugere ter 12 vezes mais itens do que a quantidade que vai ser aplicada.

#### *2.2.3.3.4 Tempo de resposta ao item e de teste*

O tempo de resposta ao item ou de teste é outra restrição que pode ser inserida em CAT. Tempo de resposta ao item (RT) é definido como o tempo decorrido entre o momento que um item é exibido e quando o respondente insere uma resposta; este tempo pode facilmente ser medido e registrado (WISE, 2014).

A aplicação de testes via computador possibilita medir com exatidão o tempo de resposta e, segundo Pasquali (2013, p. 280), o "tempo de reação vem sendo estudado desde o começo da psicologia e vem sendo considerado um elemento relevante na avaliação das aptidões humanas".

Os RTs são fontes adicionais de informação tanto em relação ao traço latente dos respondentes quanto às características do teste (HORNKE, 2000; VAN DER LINDEN, 2008). Portanto, compreender os diversos fatores que os impactam, é uma informação útil para os desenvolvedores de testes que buscam prever a quantidade de tempo necessário para aplicação completa do teste (BERGSTROM; GERSHON; LUNZ, 1994).

Em CAT, após a apresentação do item, o respondente pode ter um tempo limitado ou ilimitado para responder o item (VELDKAMP; MATTEUCCI, 2013). Quando os RTs são limitados, pode-se considerar o uso de modelos de tempo de resposta para corrigir a velocidade (*speededness*) do teste. Esta é uma informação importante quando se considera que respondentes com o mesmo nível do traço latente requerem diferentes quantidades de tempo para completar um item do teste (VAN DER LINDEN; SCRAMS; SCHNIPKE, 1999).

van der Linden, Scrams e Schnipke (1999) ressaltam que os respondentes mais lentos podem não completar todos os itens de um teste acelerado, gerando problemas por itens não respondidos; ou então, dar respostas rápidas para finalizar dentro do prazo. De acordo com Huff e Sireci (2000), quando os testes não são designados para medir "velocidade", ou seja, quando este fator não é relevante para mensurar o traço latente, ele pode ser uma potencial fonte de variância irrelevante quando este tempo é muito curto.

Para não haver problemas com a falta de tempo para responder aos itens durante o período especificado de teste, Wainer (2000) destaca que pode-se estipular um número fixo de itens, sem tempo para finalização (tempo ilimitado).

Nesse contexto, van der Linden, Scrams e Schnipke (1999) propuseram um modelo para neutralizar os diferentes efeitos dos limites de tempo em testes, beneficiando, principalmente, indivíduos com níveis de traço latente elevados. Conforme os autores, se existir um modelo para a distribuição de RTs em CAT, o tempo de resposta do indivíduo a um dado item pode ser utilizado para atualizar as estimativas destas distribuições para os itens restantes no banco. Esta distribuição de tempo de resposta, em seguida, pode ser usada para restringir a seleção de itens durante todo o teste, determinando o mesmo grau de "velocidade" para todos respondentes.

Em um estudo realizado por Chang, Plake e Ferdous (2005), conclui-se que respondentes mais capazes gastam mais tempo em todos os itens, independentemente de os itens serem respondidos corretamente ou incorretamente, de serem operacionais ou de pré-

teste, demonstrando maior persistência. Alguns estudos que analisaram a existência ou não de correlação entre o traço latente e RTs são apresentados no Quadro 6.

Quadro 6 – Resultados de estudos sobre correlação entre o traço latente e tempos de resposta.

| Autores | Resultados |
|---|---|
| van der Linden, Scrams e Schnipke (1999) | Traços latentes e "velocidade" no teste não foram correlacionados, enquanto dificuldade do item e o RT foram positivamente correlacionados. |
| Hornke (2000) | Os RTs não foram altamente correlacionados com os traços latentes, mas o tempo total de teste e a média dos RTs se correlacionaram muito bem. Além disso, indivíduos passaram mais tempo em itens que erraram do que em itens que eles acertaram. |
| Bergstrom, Gershon e Lunz (1994) | Traços latentes, geralmente, não estão relacionadas com os RTs e indivíduos com baixo nível do traço latente não demoraram mais tempo para responder aos itens do que indivíduos com nível elevado. Porém, o tamanho do enunciado do item, a posição da resposta correta e uso de figuras, contribuem para o RT. Também, indivíduos passaram mais tempo em itens que erraram do que em itens que acertaram. |
| Swanson et al. (1997) | Não houve correlação entre o traço latente e RT para prazos longos de testes, mas houve uma correlação positiva entre o traço latente e RT para prazos curtos. |

Fonte: Elaborado pela autora.

Análises de RTs no âmbito do CAT têm surgido com diferentes objetivos, como melhorar a seleção de itens quando um modelo probabilístico para a sua distribuição está disponível (VAN DER LINDEN, 2008; FAN et al., 2012; FINKELMAN et al., 2014), controlar a velocidade do teste (VAN DER LINDEN; SCRAMS; SCHNIPKE, 1999) e podem servir para ajudar a identificar comportamentos incomuns e que merecem investigação

(BERGSTROM, GERSHON; LUNZ, 1994; MEIJER; VAN KRIMPEN-STOOP, 2010; VAN DER LINDEN; VAN KRIMPEN-STOOP, 2003), como o pré-conhecimento dos itens e o acerto ao acaso (QIAN et al., 2016).

#### 2.2.3.4 Estimação do traço latente

Em CAT, três estágios de estimação podem ser destacados: (1) estimação do traço latente para iniciar o teste e, consequentemente, selecionar os itens adequados; (2) estimação do traço latente durante o teste; (3) estimação final do traço latente para informar ao respondente (VAN DER LINDEN; PASHLEY, 2010). Logo, selecionar um método adequado para este componente é especialmente importante, uma vez que isso afeta não apenas o resultado final do teste, mas também quais itens são aplicados ao longo do teste (WANG; VISPOEL, 1998).

Doebler (2012) enfatiza que a influência de efeitos nas estimativas do traço latente depende de diversos fatores, como exatidão da calibração dos itens, distribuição dos parâmetros de dificuldade do item no BI, do modelo da TRI utilizado, da duração do teste, do tamanho do BI, do método de seleção dos itens e restrições adicionais, entre outros. Deste modo, tenta-se evitar as causas desses vieses quando possível.

As principais definições para a estimação do traço latente dos respondentes em CAT são apresentadas a seguir, e foram retiradas de van der Linden e Pashley (2010). Considere $q$ como sendo o item de interesse, mas não previamente aplicado, de um BI com $i = 1, ..., I$ itens; o *rank* dos itens em um CAT é denotado por $k = 1,..., K$, ou seja, é o número de itens aplicados; logo, $i_k$ é o índice do item no BI aplicado como o $k$-th item no teste para um respondente.

A seguir, considere a seleção do $k$-th item no teste. Os itens $k$-1 anteriores formam um conjunto $S_k = \{i_i,...,i_{k-1}\}$. Eles têm respostas que são representadas por realizações das variáveis resposta $U_{i_1} = u_{i_1},..., U_{i_{k-1}} = u_{i_{k-1}}$. O conjunto de itens restante no BI após $k$-1 itens serem selecionados é $R_k = \{1, ..., I\} \setminus S_{k-1.}$ O item $k$ é selecionado a partir deste conjunto.

A função de verossimilhança associada com as respostas nos primeiros *k* -1 itens é:

$$L\left(\theta \mid u_{i_1} \ldots u_{i_{k-1}}\right) \equiv \prod_{q=1}^{k-1} \frac{\left\{\exp\left[a_{i_q}\left(\theta - b_{i_q}\right)\right]\right\}^{u_{i_q}}}{1+\exp\left[a_{i_q}\left(\theta - b_{i_q}\right)\right]} \tag{2}$$

A derivada de segunda ordem da log-verossimilhança reflete na curvatura da função de verossimilhança observada no $\theta$ relativo à escala para este parâmetro. O negativo dessa derivada é geralmente conhecido como a medida de informação observada:

$$J_{u_{i_1} \ldots\, u_{i_{k-1}}}(\theta) \equiv -\frac{\partial}{\partial \theta^2} \ln L\left(\theta \mid u_{i_1}, \ldots,\, u_{i_{k-1}}\right) \tag{3}$$

O valor esperado da medida de informação observada sobre as variáveis resposta é a medida de informação esperada de Fisher:

$$I_{u_{i_1} \ldots\, u_{i_{k-1}}}(\theta) \equiv E\left[J_{u_{i\,1} \ldots\, u_{i\,k-1}}(\theta)\right] \tag{4}$$

Para o modelo descrito na Equação 1, a medida de informação esperada se reduz a:

$$I_{u_{i_1} \ldots\, u_{i_{k-1}}}(\theta) = \sum_{q=1}^{k-1} \frac{[p'_{i_q}(\theta)]^2}{p_{i_q}(\theta)[1-p_{i_q}(\theta)]} \tag{5}$$

com $p'_{i_q}(\theta) \equiv \frac{\partial}{\partial \theta} p_{i_q}(\theta)$, onde $p_{i_q}(\theta)$ é a probabilidade de responder corretamente ao item de interesse *q* ($i_q$) dado o traço latente.

Em CAT, tanto a função de informação de Fisher quanto a função de informação observada podem ser utilizadas no algoritmo para seleção do próximo item quando o método MEI é utilizado (VAN DER LINDEN, 1998).

Na abordagem Bayesiana, assume-se uma distribuição *a priori* para os valores desconhecidos dos parâmetros do traço latente, g($\theta$), que junto com a verossimilhança, formam a distribuição *a posteriori* de $\theta$.

$$g(\theta|_{u_{i_1}\ldots\ u_{i_{k-1}}}) = \frac{L(\theta \mid u_{i_1},\ldots,\ u_{i_{k-1}})g(\theta)}{\int L(\theta \mid u_{i_1},\ldots,\ u_{i_{k-1}})g(\theta)d\theta} \tag{6}$$

Tipicamente, essa densidade é assumida ser uniforme ou como uma estimativa empírica da distribuição do traço latente da população de respondentes, isto é, uma distribuição *a priori* normal.

O estimador do traço latente após as respostas a *k*-1 itens é denotado como $\hat{\theta}_{u_{i_1}\ldots u_{i_{k-1}}} = \hat{\theta}_{k-1}$. O estimador de máxima verossimilhança visa maximizar a função de verossimilhança dada na Equação 2 sobre possíveis valores de $\theta$.

$$\hat{\theta}^{MV}_{u_{i_1}\ldots u_{i_{k-1}}} \equiv \arg\max_{\theta}\{L(\theta \mid u_{i_1}\ldots u_{i_{k-1}}) : \theta \in (-\infty,\infty) \tag{7}$$

Uma alternativa é o estimador ponderado pela verossimilhança (WLE) de Warm (1989), o qual maximiza a verossimilhança dada na Equação 2 ponderada por uma função $w_{k-1}(\theta)$:

$$\hat{\theta}^{WLE}_{u_{i_1}\ldots u_{i_{k-1}}} \equiv \arg\max_{\theta}\{w_{k-1}(\theta)L(\theta \mid u_{i_1}\ldots u_{i_{k-1}}) : \theta \in (-\infty,\infty) \tag{8}$$

onde a função $w_{k-1}(\theta)$ é definida para satisfazer:

$$\frac{\partial w_{k-1}(\theta)}{\partial \theta^2} \equiv \frac{H_{k-1}(\theta)}{2I_{k-1}(\theta)} \tag{9}$$

com $H_{k-1}(\theta) \equiv \sum_{q=1}^{k-1} \frac{[p'_{i_q}(\theta)][p''_{i_q}(\theta)]}{p_{i_q}(\theta)[1-p_{iq}(\theta)]}$ e $p''_{i_q}(\theta) \equiv \frac{\partial^2 p_{i_q}(\theta)}{\partial \theta^2}$,

e $I_{k-1}(\theta) \equiv I_{U_{i_1} \ldots U_{i_{k-1}}}(\theta)$ como definido na Equação 4.

Na abordagem Bayesiana, um estimador pontual de $\theta$ pode ser baseado na sua distribuição *a posteriori* dada na Equação 6. Os estimadores são MAP e EAP.

$$\hat{\theta}^{MAP}_{u_{i_1} \ldots u_{i_{k-1}}} \equiv \arg\max_{\theta} \{g(\theta \mid u_{i_1} \ldots u_{i_{k-1}}) : \theta \in (-\infty, \infty) \quad (10)$$

O estimador EAP é o valor esperado da distribuição *a posteriori*,

$$\hat{\theta}^{EAP}_{u_{i_1} \ldots u_{i_{k-1}}} \equiv \int \theta \, g(\theta \mid u_{i_1} \ldots u_{i_{k-1}}) d\theta \quad (11)$$

Outra possível abordagem é abster-se de estimativas pontuais e usar a distribuição *a posteriori* completa de $\theta$ como o estimador. Este estimador não só revela o valor mais plausível de $\theta$, mas mostra também a plausibilidade de qualquer outro valor. É comum resumir esta incerteza sobre $\theta$ sob a forma de variância da distribuição *a posteriori* de $\theta$, dado por

$$\mathrm{Var}(\theta \mid u_{i_1} \ldots u_{i_{k-1}}) \equiv \int [\theta - E(\theta \mid u_{i_1} \ldots u_{i_{k-1}})]^2 \, g(\theta \mid u_{i_1} \ldots u_{i_{k-1}}) d\theta \quad (12)$$

De acordo com van der Linden e Pashley (2010), quando ocorrem padrões de resposta com todos os itens corretos ou incorretos, não existem estimativas MV finitas e, em CAT, as propriedades de amostras pequenas do estimador MV dependerão de fatores como a distribuição dos itens no BI e do método utilizado para a seleção dos itens. Devido ao fato de que esses padrões podem ocorrer na apresentação dos itens iniciais em CAT, o método MV não pode ser utilizado, a menos que algumas estratégias sejam utilizadas para tentar resolver este problema, como apresentar alguns itens iniciais e posteriormente estimar o traço latente.

Este problema gerado pelo método MV faz com que métodos Bayesianos sejam preferidos. No entanto, conforme van

der Linden e Pashley (2010), eles também dependem da distribuição *a priori* utilizada, que pode influenciar nos itens iniciais selecionados para aplicação. Segundo os autores, más estimativas iniciais podem interferir apenas em testes curtos, por exemplo, com 10 itens. Para os testes mais longos, entre 20 a 30 itens, não terá impacto.

Para uma distribuição *a priori* uniforme na Equação 6, o máximo das Equações 7 e 9 são iguais. Neste caso, MAP tem as mesmas propriedades de MV. Para distribuições *a priori* não-uniformes, as propriedades de amostras pequenas do estimador MAP dependem não só da verossimilhança, mas também da distribuição *a priori*, que dependendo da escolha, a distribuição *a posteriori* pode ser multimodal, resultando em um máximo local, se precauções não forem tomadas. Por outro lado, se uma distribuição *a priori* adequada for usada, o estimador EAP sempre existirá e ao contrário dos outros estimadores, é fácil de calcular (VAN DER LINDEN; PASHLEY, 2010).

Por fim, é importante destacar que nenhuma destas soluções é inteiramente satisfatória, por isso a importância das simulações para auxiliarem na tomada de decisão do melhor método a ser utilizado para cada situação particular. As avaliações dos métodos são normalmente baseadas em índices como RMSE, viés e EP (WANG; VISPOEL, 1998; WANG; HANSON; LAU, 1999).

O viés causa um impacto negativo sobre a validade de um teste, podendo tornar mais difícil manter a comparabilidade entre CAT e outras versões do teste, como em P&P (WANG; HANSON; LAU, 1999). De acordo com Wang e Vispoel (1998), a escolha entre MV e uma abordagem bayesiana dependerá do papel que o EP e viés exercem na tomada de decisões a partir dos resultados do CAT. Assim, tem-se (WANG; VISPOEL, 1998; WANG; HANSON; LAU, 1999):

- Se o objetivo é a ordem de classificação dos respondentes, então EP será de maior importância do que o viés porque não vai afetar a ordem das estimativas do traço latente. Sendo assim, métodos bayesianos fornecem melhores resultados, especialmente em testes de comprimento menores de 30 itens; acima de 30 itens, há poucas

diferenças significativas entre os métodos. Doebler (2012) também ressalta que o estimador pode influenciar apenas em testes curtos.

- Se o objetivo de um CAT é comparar médias de grupos, referenciar estimativas do traço latente de diferentes testes na mesma escala (por exemplo, CAT e P&P) ou decisões de aprovado/reprovado, certificações ou licenciamento, o viés será mais importante do que EP. Nesses casos, o viés pode causar mudanças sistemáticas nas médias dos grupos, nas estimativas individuais do traço latente e em pontos de corte de classificação, especialmente quando os níveis extremos do traço latente estão envolvidos. Em tais situações, MV seria preferível, a não ser que uma distribuição *a priori* possa ser ajustada para eliminar o problema de viés.

O Quadro 7 apresenta alguns resultados de estudos que comparam métodos de estimação do traço latente em CAT.

#### 2.2.3.5 Regra de finalização do teste

A regra de parada ou finalização de um CAT determina quando a aplicação de itens deve acabar, podendo, o teste, ter comprimento fixo ou variável. CATs de comprimento fixo aplicam a mesma quantidade de itens para cada respondente, enquanto CATs de comprimento variável terminam quando certo critério predeterminado de precisão de medida é atingido (YI; WANG; BAN, 2001).

Outra possível regra de parada é definir um limite de tempo para todo o teste. Este critério geralmente é combinado com CAT de comprimento fixo ou variável (VELDKAMP; MATTEUCCI, 2013). No entanto, apesar dos motivos práticos para impor limite de tempo, Matteucci e Veldkamp (2013) destacam que este critério deve ser usado com cautela, uma vez que pode ameaçar a validade do teste quando o teste é acelerado, pois poderá não medir efetivamente o traço latente investigado.

Quadro 7 – Comparação de métodos de estimação do traço latente em CAT.

| Autores | Métodos clássicos | Métodos bayesianos | Resultados |
|---|---|---|---|
| Wang (1997) | MV | EU-EAP (EAP com distribuição *a priori* Beta) | EU-EAP pode produzir estimativas com viés semelhante ou até menor do que o MV e sem sacrificar muito o EP e RMSE, como ocorre na estimativa EAP padrão. As restrições práticas como BC e controle de taxa de exposição do item não afetam o viés para EU-EAP. |
| Wang e Vispoel (1998) | MV | EAP e MAP | Comparam métodos sob diferentes cenários CAT (variando as características do BI, regras de inicialização e finalização do teste - sem taxa de exposição e BC). MV produziu menor viés, maiores EP e RMSE, menor fidelidade e eficiência na aplicação, sendo os bayesianos mais adequados; EAP proporcionou o melhor resultado. |
| Wang, Hanson e Lau (1999) | MV e WLE | MAP e EU-MAP (MAP com distribuição *a priori* Beta) e EU-EAP | Compararam os efeitos que a forma da distribuição *a priori*, diferentes características do BI e restrições práticas (BC e taxa de exposição) têm sob o viés, EP e RMSE. No geral, EU-MAP apresentou melhor desempenho, reduzindo significativamente o viés em testes de comprimento fixo (embora com um ligeiro aumento no RMSE) e desempenhou razoavelmente bem quando foi usada uma regra de parada com base na variância posterior predeterminada. MV apresentou os piores resultados. |
| Cheng e Liou (2000) | MV e WLE | __ | Os autores também utilizaram três diferentes métodos de seleção de itens e concluíram que a maioria dos algoritmos CAT desempenhou igualmente bem para testes com 10 itens ou mais. |

Fonte: Elaborado pela autora.

De acordo com Matteucci e Veldkamp (2013), CAT de comprimento fixo é muitas vezes aplicado quando o teste tem que cumprir uma série de especificações com relação à conteúdo ou outros atributos; já o teste de comprimento variável é comumente adotado quando se deseja garantir que todos os respondentes sejam medidos com o mesmo nível de precisão, embora alguns indivíduos possam ter de responder mais itens do que outros.

Porém, se o EP é fixo, deve-se ficar atento aos níveis de dificuldade dos itens que compõem o BI, principalmente para respondentes com traço latente muito baixo ou muito elevado (YI; WANG; BAN, 2001; MATTEUCCI; VELDKAMP, 2013). Nesses casos, geralmente há uma combinação do EP com um número máximo de itens. O número mínimo de itens para atingir a precisão aceitável é definido por simulação antes da aplicação efetiva de um CAT.

Outras formas de finalizar o teste estão relacionadas à exaustão do BI ou à pequenas mudanças em critérios predeterminados, tais como: pequenas diferenças nas estimativas do traço latente; por meio da informação mínima, ou seja, quando não há mais itens restantes no BI que forneçam uma quantidade predeterminada de informação para o respondente (LEROUX et al., 2013); quando a redução de erros padrão sucessivos é menor ou igual do que um valor definido. A desaceleração dessa redução no EP indica que os itens que restam no BI são pouco informativos para o traço latente do respondente (NUNES et. al, 2015).

Em aplicações convencionais como de certificação, nivelamento e classificação, costuma-se utilizar um número fixo de itens (MILLS; STOCKING, 1995). Nestes casos, comprimento variável pode não ser viável, quer porque o conteúdo do teste é especificado em pormenor ou porque respondentes podem perceber o teste como injusto (VELDKAMP; MATTEUCCI, 2013).

Deste modo, a escolha de uma regra de finalização do teste depende do contexto do teste e de fatores como o tempo de teste, eficiência da medição, da comparabilidade com outras formas de aplicação de testes, entre outros fatores (YI; WANG; BAN, 2001). A desvantagem de testes com comprimento fixo é que os

respondentes terão diferentes precisões de medida do traço latente dependendo do nível em que ele se encontra na escala (MILLS; STOCKING, 1995).

## 2.3 *SOFTWARES* E PLATAFORMAS PARA CAT

Um *software* completo para avaliação adaptativa deve conter vários módulos para contemplar as diferentes etapas que o compõem. Assim, conforme Hambleton, Zaal e Pieters (1991), o *software* deve ter:

- Procedimentos de identificação dos respondentes e dos testes aplicados;
- Textos e parâmetros do banco de itens;
- Um módulo de construção de testes;
- Um módulo de apresentação de itens (início do teste, seleção de itens, finalização do teste, estimação final do traço latente e precisão) e armazenamento dos resultados;
- Um módulo de atualização do banco de itens (desempenho dos respondentes e informação histórica dos itens);
- Um módulo para oferecer ao usuário uma informação detalhada do seu desempenho.

Os autores Bergstrom e Gershon (1995) apresentam questões práticas para a construção de um BI computadorizado, destacando quais informações dos itens que devem ser armazenadas e questões para maximizar a eficiência computacional. Assim, depois de elaborado o *software*, ele deverá ser implementado em computadores, servidores ou na internet, conforme o propósito do teste (MOREIRA JUNIOR, 2011).

O Quadro 8 apresenta algumas plataformas que dão suporte para o desenvolvimento e administração do CAT. Há, também, diversos *softwares* disponíveis para efetuar simulações em CAT, a fim de auxiliar na definição do *design* do teste. O Quadro 9 apresenta uma lista de pacotes do *software* R que podem ser utilizados para este fim.

Para modelos unidimensionais, o pacote catR parece ser o mais completo em opções implementadas para o algoritmo. Ele também pode ser usado em conjunto com o Concerto para a aplicação de testes aos respondentes. O Quadro 10 apresenta uma lista de outros *softwares* disponíveis que também foram desenvolvidos para simulações de CATs.

Quadro 8 – Plataformas disponíveis para administrar o CAT.

| **Plataforma** | **Objetivo** | **Desenvolvimento** | **Licença** | ***Link*** |
|---|---|---|---|---|
| Assessment Center | Sistema para desenvolvimento e administração do CAT | Assessment Center$^{SM}$ - Northwestern University | livre | http://www.assessmentcenter.net/ |
| Cito | Sistema para desenvolvimento e administração de testes | Cito | comercial | http://www.cito.com/research_and_development/computer_webbased_testing.aspx |
| Concerto | Sistema para desenvolvimento e administração de testes | Universidade de Cambridge | Livre | https://code.google.com/p/concerto-platform/ |
| eTests | Administração de testes | CASAS | comercial | https://www.casas.org/product-overviews/software/casas-etests |
| ETS e suas subsidiarias (ex.: CA&L e Prometric) | Sistema para desenvolvimento e administração de testes | Educational Testing Service | comercial | https://www.ets.org/about |
| FastCAT (antigo FastTEST) | Sistema para desenvolvimento e administração de testes e análise psicométrica | Assessment Systems Corporation | comercial | http://www.assess.com/xcart/ |
| Pearson VUE | Sistema para desenvolvimento e administração de testes | Pearson VUE | comercial | http://www.pearsonvue.com/ |

Continuação

| Plataforma | Objetivo | Desenvolvimento | Licença | *Link* |
|---|---|---|---|---|
| IRT-Computerized Adaptive Testing | Programa de acesso aberto com base na internet para CAT. Adiciona análise de itens com base na TRI para os modelos ML3P e Rasch. | | livre | http://sourceforge.net/projects/irt-cat/ |
| MATE | Simulação e administração de CAT multidimensional. | DIPF TBA - Technology Based Assessment | livre | http://tba.dipf.de/en/assessment/mate-1 |
| McCann | Várias plataformas que dão suporte para a implantação de testes para fins profissionais e educacionais. | McCann Associates | comercial | http://www.mccanntesting.com/ |
| mirtCAT | Fornece ferramentas para gerar uma *interface* HTML para criar CATs unidimensionais e multidimensionais. Usado junto com o pacote Shiny (CHANG et al., 2015). | pacote do R (CHALMERS, 2015) | livre | https://cran.r-project.org/web/packages/mirtCAT/index.html |

Continuação

| Plataforma | Objetivo | Desenvolvimento | Licença | *Link* |
|---|---|---|---|---|
| OSCATS | Biblioteca para os modelos psicométricos e algoritmos utilizados em CAT - em simulações ou como parte de um CAT operacional. | | livre | https://code.google.com/p/oscats/ |
| SIETTE | Sistema para desenvolvimento e administração de testes. | Malaga University | livre | http://portal.siette.org/index.php?lang=es |
| Smart Test Technology® | Plataforma para administração do CAT | Adaptive Assessment Services, Inc | comercial | http://aastest.com/ |
| SmarterApp | Sistema para desenvolvimento e administração de testes. | Smarter Balanced Assessment System | livre | http://www.smarterbalanced.org/ ou http://www.smarterapp.org/ |
| WebeXaminer FASTCAT | Sistema para desenvolvimento e administração do CAT e análise psicométrica. | WebExaminer | comercial | http://www.webexaminer.com/computer_adaptive.php |

Fonte: Elaborado pela autora.

Quadro 9 – Pacotes do *software* R para simulações CAT.

| Pacote do R | Modelo | Regras |
|---|---|---|
| mirtCAT (CHALMERS, 2015) | Modelos unidimensionais e multidimensionais | Fornece ferramentas para gerar uma *interface* HTML para a criação de CAT unidimensionais e multidimensionais. Deve ser utilizado junto com o pacote Shiny (CHANG et al., 2015). |
| catIrt (NYDICK, 2014) | Modelos dicotômicos e politômicos | 1) Método para estimar o traço latente: MV, MAP, EAP, WLE; 2) seleção do próximo item: MFI, MLWI, MPWI, *random*; 3) regra de parada: comprimento, precisão e classificação (por *Confidence interval method*, *Sequential Probability Ratio Test*, *Generalized Likelihood Ratio*); 4) controle da taxa de exposição: método SH. |
| catR (MAGIS; RAICHE, 2012) | Modelos dicotômicos e politômicos | 1) Método para estimar o traço latente: MV, MAP, EAP, WLE; 2) seleção do próximo item: MFI, MLWI, MPWI, MEI, MEPV, KL, KLP, PG, PP, *random*; 3) regra de parada: comprimento, precisão e classificação (por *Confidence interval method*); 4) controle da taxa de exposição: métodos *randomesque*, *Progressive* e *proportional*; 5) possui restrição para BC. |
| MAT (CHOI; KING, 2014) | Modelo de três parâmetros multidimensional | 1) Método para estimar o traço latente: MAP e o método *Scoring* de Fisher; 2) seleção do próximo item: D-*optimality*, A-*optimality*, e C-*optimality;* 3) regra de parada: pode ser especificado como um critério conjuntivo ou um critério de compensação; 4) controle de exposição dos itens: *randomesque*; 5) possui restrição para BC, especificando distribuições de conteúdo alvo. |

Fonte: Elaborado pela autora.

Quadro 10 – Outros *softwares* para simulações CAT.

| ***Software*** | **Função** | ***Link*** | **Licença** | **Algoritmo** |
|---|---|---|---|---|
| CATSim (antigo PostSim) | Simulação de CAT para itens dicotômicos e politômicos. | www.assess.com/xcart | Comercial | Estimação do traço latente por MV, WLE, EAP, MAP; três métodos de seleção de itens; regra de parada: EQM ou informação; com BC; controle de exposição dos itens; itens inimigos. |
| Firestar | Simulação de CAT para itens politômicos. | www.nihpromis.org/resources/resourcehome | Livre | Estimação do traço latente por EAP, MAP, WLE, MV; nove métodos de seleção de itens; regra de parada: comprimento fixo e variável; controle da exposição. |
| Firestar-D | Simulação de CAT para itens dicotômicos. | Solicitar ao autor S.W.Choi s-choi@northwestern.edu | Livre | Não apresenta. |
| SIMPOLYCAT (macro do SAS) | Simulação de CAT para itens politômicos. | www.sas.com/pt_br/home.html | Comercial | Estimação do traço latente por MV, MAP, EAP, WLE; dois métodos de seleção de itens; regra de parada: comprimento fixo e variável; controle da exposição. Não tem BC, mas permite usar subconjunto de itens para dependência local. |
| SimulCAT | *Software* para simular administração de CAT | www.umass.edu/remp/software/simcata/simulcat/ | Livre | Estimação do traço latente por MV, MAP, EAP; sete métodos de seleção de itens; controle da exposição do item; comprimento fixo e variável. Permite informar a existência de DIF, DRIFT e dados de exposição dos itens preexistentes. |
| SimuMCAT | Simulação de CAT multidimensional (ML3P e de crédito parcial) | www.bmirt.com/6271.html | Livre | Estimação do traço latente por MAP e MV; cinco métodos de seleção de itens; controle da exposição; BC; comprimento fixo e variável. |

Fonte: Elaborado pela autora.

## 3. SISTEMÁTICA PARA MANUTENÇÃO DO BANCO DE ITENS

O uso da tecnologia em conjunto com o desenvolvimento de um BI pode facilitar tarefas relacionadas à elaboração dos testes e melhorar a qualidade dos itens que o compõem e, consequentemente, das avaliações (WARD; MURRAY-WARD, 1994). A praticidade que um BI informatizado oferece, pode trazer benefícios tanto para grandes programas de avaliações quanto para organizações menores (BERGSTROM; GERSHON, 1995).

Embora a troca para um BI informatizado envolva custos iniciais com o desenvolvimento ou aquisição de um *software* e com possível conversão de itens existentes, gráficos e estatísticas para este novo modo, reduções de custos são obtidos em longo prazo com a redução de tempo profissional e administrativo para a elaboração dos testes e redução de erros por manipulação manual de dados (BERGSTROM; GERSHON, 1995).

Para um BI cumprir com sua função, além de ser bem elaborado, ele necessita de manutenção (BERGSTROM; GERSHON, 1995, SQUIRES, 2003). Pasquali (2013, p. 282) reitera que "é preciso fazer a manutenção periódica dos bancos de itens, porque, com o tempo, eles podem perder suas características de bons itens, isto é, sua validade". No entanto, a manutenção do BI e a segurança dos testes é um desafio, principalmente quando o teste é aplicado com certa frequência. Desta forma, serão necessários procedimentos para acompanhar o funcionamento do BI e seus itens ao longo do tempo.

O trabalho de manutenção de um BI geralmente é feito por profissionais especialistas de conteúdo e psicometristas (BERGSTROM; GERSHON,1995; SQUIRES, 2003). Psicometria é a área de estudo que utiliza conceitos especializados e métodos matemáticos e estatísticos para estabelecer a confiabilidade e validade de um teste (SQUIRES, 2003).

De acordo com Bergstrom e Gershon (1995), especialistas de conteúdo devem revisar sistematicamente o BI para assegurar que: (1) os itens são atuais e relevantes para o campo da prática; (2) itens duplicados e semelhantes são identificados e sinalizados; e (3) o conteúdo dentro do BI é representativo do construto a ser mensurado.

Psicometristas também devem revisar o BI para garantir que: (1) existam itens com diferentes níveis de dificuldade e com boa capacidade de discriminação, cobrindo todos os pontos da escala; e (2) a aprovação/reprovação dos itens é atual (BERGSTROM; GERSHON, 1995).

Way (2006) reforça a importância da manutenção de um BI para CAT, uma vez que itens são aplicados muitas vezes durante anos, e esta prática levanta questões como:

- Itens superexpostos devem ser retirados do uso para não comprometer a avaliação;
- Itens podem se tornar desatualizados, especialmente em áreas das ciências e estudos sociais. Logo, procedimentos formais para rever e confirmar se os itens continuarão a ser elegíveis para o uso ao longo dos anos deve ser introduzido;
- Muitos programas de avaliação exigem a divulgação dos itens ao público após a aplicação. Essa política de divulgação tornaria muito cara a manutenção do BI, exigindo esforços significativos no desenvolvimento de itens para repor continuamente os BIs divulgados, dado que cada respondente recebe uma forma diferente de teste.

Conforme ressaltam Mills e Stocking (1995), esta política provavelmente resultaria na divulgação de itens mais rápido do que eles podem ser desenvolvidos e pré-testados. Neste caso, uma forma de proteger o BI é não divulgar ou disponibilizar aos respondentes os itens do teste, nem o gabarito. Ao final do teste, o respondente recebe um *feedback* do seu traço latente juntamente com um relatório contendo os conhecimentos ou habilidades que ele domina.

Na tentativa de reduzir custos com o desenvolvimento de itens, Glas e van der Linden (2003) propuseram a utilização de técnicas de geração de itens baseada em regra (clonagem de itens) para aumentar o número de itens disponíveis no BI. A partir da descrição formal de um conjunto de "itens pai", ou seja, formas, modelos de itens e algoritmos para derivar famílias de clones a partir deles, geram-se itens. Por exemplo, descreve-se a sintática de um item com um ou mais locais abertos para os quais conjuntos para substituição são especificados (distratores são escolhidos aleatoriamente e inseridos nos locais abertos). Porém, esta técnica pode tornar as características estatísticas dos itens de uma família clonada incomparáveis. Desta forma, os autores testam métodos para contornar este problema, os quais mostraram que podem reduzir a precisão do teste.

O principal índice de segurança dos itens e monitoramento tradicionalmente utilizado é a taxa de exposição dos itens; por isso, vários algoritmos de controle foram desenvolvidos (WANG; KOLEN, 2001). No entanto, fica evidente que somente o controle da taxa de exposição

não basta para a manutenção, assim como apenas a calibração de novos itens é insuficiente.

Portanto, propõem-se a sistemática apresentada a seguir, que aborda um conjunto de ações para a efetiva manutenção do BI para testes de alto impacto, com determinada sequência de procedimentos a serem adotados, combinados à disponibilidade de um grande BI com boa qualidade psicométrica. Para operacionalizar cada etapa da manutenção, são apresentados os métodos disponíveis especificamente para o contexto de CATs com base em estudos já realizados.

Em testes de alto impacto, não é esperado alterar o algoritmo do CAT devido às deficiências no BI. Por isso essas decisões precisam ser bem planejadas e estudadas previamente na definição do *design* do CAT.

## 3.1 PROCEDIMENTOS PROPOSTOS PARA A MANUTENÇÃO DO BI PARA CAT DE ALTO IMPACTO

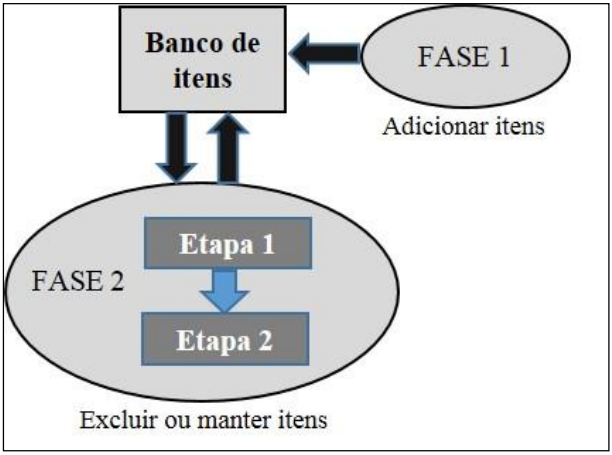
No desenvolvimento da sistemática para manutenção do BI de CAT aplicados em avaliações de alto impacto, optou-se por criar uma metodologia dividida em duas fases. Todas as etapas estão relacionadas aos itens de um BI e o ideal é que todas elas sejam seguidas em cada "edição" de aplicação dos testes, seja antes de iniciar (Fase 1) ou após sua conclusão para a análise dos itens que já estão no BI (Fase 2). Assim, de forma geral, tem-se:

- **Fase 1: CALIBRAÇÃO DE NOVOS ITENS** - envolve a definição de quantos e quais itens serão calibrados; definição do *design* de calibração, ou seja, do método de seleção de itens de pré-teste, do local de inserção do item de pré-teste no CAT, do método de estimação dos parâmetros do item e da regra de parada para itens de pré-teste; além da análise dos dados para a inclusão efetiva desses itens no BI.
- **Fase 2: MONITORAMENTO DOS ITENS** - esta fase aborda duas etapas: **Etapa 1:** MONITORAMENTO DA EXPOSIÇÃO DOS ITENS - envolve a análise das taxas de exposição dos itens e sobreposição de testes e identificação de itens pré-conhecidos; **Etapa 2:** VERIFICAÇÃO DE DRIFT DOS PARÂMETROS DOS ITENS - envolve técnicas que visam identificar se itens tiveram seus parâmetros alterados ao longo do tempo.

Essas fases podem ser representadas conforme a Figura 6. A fase 1 visa adicionar novos itens ao BI, na tentativa de ampliá-lo. Já a fase 2

visa monitorar o item por meio das etapas 1 e 2 e tomar a decisão de mantê-lo ou excluí-lo do BI. Caso nenhuma inconsistência seja encontrada, o item retorna para o BI para ser novamente utilizado em CATs futuros.

Figura 6 – Fases do processo de manutenção do BI.




Fonte: Elaborado pela autora.

Para iniciar a implementação da sistemática, considere inicialmente um BI calibrado pela TRI e um algoritmo CAT que incorpore a restrição de controle da taxa máxima de exposição dos itens no algoritmo de seleção dos itens. A restrição de balanceamento de conteúdo pode ou não ser inserida, dependendo do traço latente que está sendo investigado. No entanto, como ela é uma restrição muito importante para mensurar adequadamente a maioria dos traços latentes da natureza, sua inserção será considerada para a sistemática.

Os procedimentos metodológicos para operacionalizar cada fase da sistemática são apresentados de forma mais genérica nas figuras e, posteriormente, são detalhados no texto.

### 3.1.1 FASE 1 - Calibração de novos itens

Novos itens devem ser escritos, pré-testados e adicionados ao BI para a manutenção (STOCKING, 1994; SQUIRES, 2003). Assim, esta etapa de calibração de novos itens visa cobrir lacunas de conteúdo e melhorar as propriedades psicométricas e o desempenho de um CAT, bem como repor itens que se tornam obsoletos ou superexpostos com o tempo

e precisam ser eliminados do BI (BAN et al., 2001; VELDKAMP; MATTEUCCI, 2013; ZHENG, 2014). A eliminação de itens nestas situações será discutida na Fase 2 da sistemática.

Quando um item está pronto para ser utilizado no traço latente dos respondentes, ele é referido como um item operacional; antes disso, ele é referido como um item de pré-teste, pois ainda precisa passar por etapas de avaliação (NANDAKUMAR; ROUSSOS, 2004). Normalmente, esses itens de pré-teste não são incluídos na estimativa do traço latente do respondente; somente depois de calibrados poderão ser usados para este fim (BOCK; MURAKI; PFEIFFENBERGER, 1988; SQUIRES, 2003; ZHENG, 2014).

Stocking (1994) sugere duas formas para calibração de itens de pré-teste para serem inseridos no BI:

- Teste aplicado via P&P, contanto que seja demonstrado que os parâmetros dos itens não mudam quando muda o modo de aplicação (ver GREEN et al., 1984; WANG; KOLEN, 2001; GWALTNEY; SHIELDS; SHIFFMAN, 2008; ZITNY et al., 2012; RILEY; CARLE, 2012);
- Itens de pré-teste (novos) são aplicados junto com os itens operacionais durante o CAT para a obtenção de estimativas dos parâmetros dos itens. Este método é denominado de calibração on-line (STOKING, 1988b; GUO; WANG, 2003; THOMPSON; WEISS, 2011; MASTERS; MUCKLE; BONTEMPO, 2009; ZHENG, 2014) e será discutido ao longo desta seção.

É importante destacar que, quanto mais semelhante com o teste adaptativo for o método de aplicação dos itens de pré-teste para calibração, menores serão os possíveis efeitos nas estimativas dos parâmetros dos itens (WANG; KOLEN, 2001; BJORNER et al., 2007; DAVEY, 2011; VAN DER LINDEN; REN, 2015; MAKRANSKY; GLAS, 2010). Por isso, considera-se para esta sistemática, que os itens são aplicados junto ao CAT operacional, visando minimizar esses efeitos e aproveitando a própria aplicação do teste para calibrar novos itens, sem custos adicionais de coleta de dados com outra aplicação.

A aplicação junto ao CAT operacional também permite obter amostras grandes, representativas e motivadas dos respondentes em uma situação real de avaliação (DAVEY, 2011; ALI; CHANG, 2014; VAN DER LINDEN; REN, 2015; ZHENG, 2014), uma vez que o respondente não identifica quais são os itens novos que estão sendo calibrados (ALI; CHANG, 2014).

A Figura 7 apresenta os procedimentos para esta primeira fase de manutenção do BI, a qual envolve diferentes aspectos relacionados à expansão do BI. Para dar início a esta fase, é necessário que itens de pré-teste estejam disponíveis para serem aplicados. Caso contrário, novos itens devem ser elaborados com urgência e esta etapa de manutenção não é executada junto ao CAT operacional, o qual é aplicado normalmente, conforme regras definidas.

Supondo a existência de itens para serem pré-testados, os especialistas devem definir quais e quantos itens devem ser pré-testados e qual será o *design* de calibração on-line a ser adotado. O CAT é então aplicado e os dados são armazenados. Uma vez que o item de pré-teste atinge a regra de parada predefinida e tem seus parâmetros estimados, é preciso avaliar a qualidade do item. Caso o item não atinja a regra de parada, ele é novamente aplicado.

Assim, se o item apresenta características adequadas, ele pode ser inserido no BI e a função de informação do BI (FIBI) é atualizada para fornecer uma visão geral sobre o BI atual, finalizando a Fase 1. Se o item apresentar problema(s) nesta fase, deve ser enviado para análise por especialistas de conteúdo e, portanto, não vai para o BI, finalizando a Fase 1.

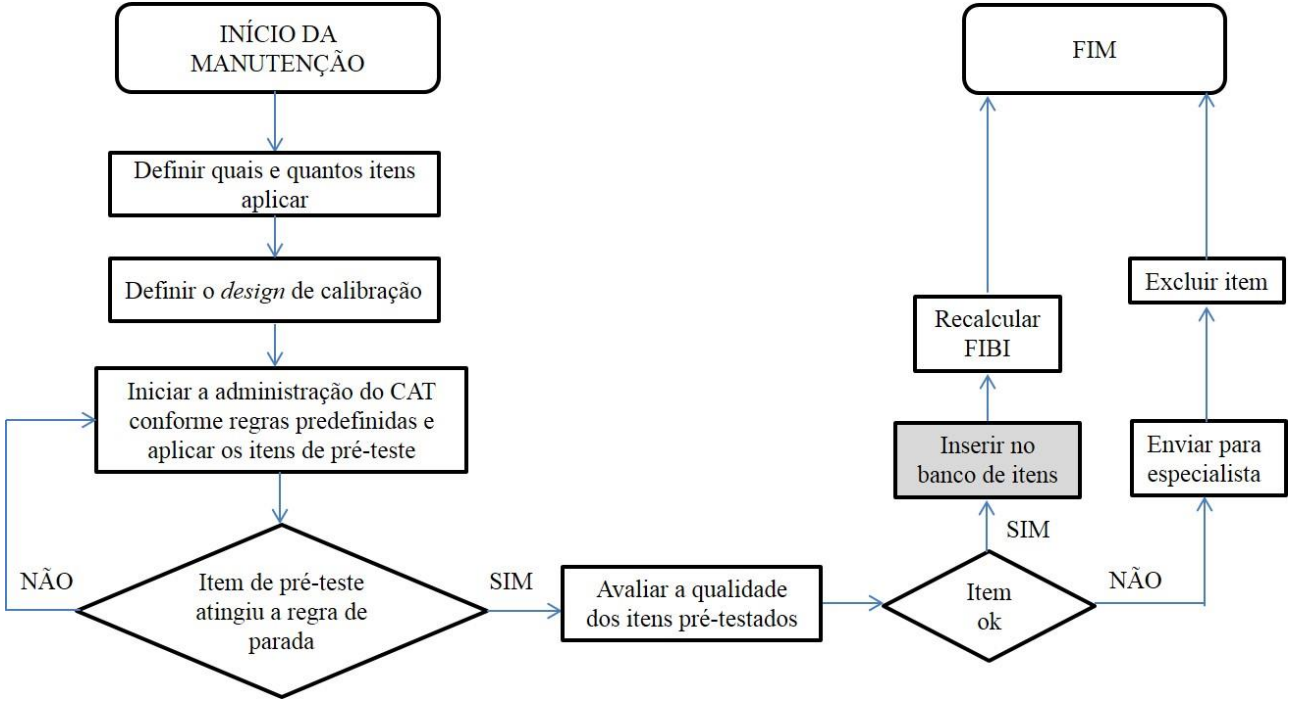
Os procedimentos para dar suporte a essas ações são detalhados a seguir. Para iniciar este processo de manutenção, duas questões precisam ser respondidas:

- **Quantos itens de pré-teste aplicar e qual o tamanho da amostra?**

Não há um número máximo ou mínimo de itens de pré-teste que podem ser aplicados a cada edição do teste. A ideia é de quanto mais itens no BI, melhor para estimar com precisão o traço latente e reduzir a taxa de exposição dos itens. Porém, a definição do número de itens de pré-teste que serão aplicados depende do número de respondentes do teste. Se muitos itens de pré-teste estão disponíveis e a amostra for pequena, a aplicação de inúmeros itens de pré-teste poderá levar à imprecisão das estimativas dos parâmetros desses itens.

O comprimento do teste deve ser considerado nesta decisão de quantos itens serão pré-testados, uma vez que pode causar um impacto negativo no processo de calibração se ele for muito extenso. Como regra geral, Ali e Chang (2014) sugerem que o número de itens de pré-teste fique entre 25% e 33% do comprimento do CAT operacional.

Figura 7 – Representação da FASE 1 de manutenção do BI: Calibração de novos itens.




Fonte: Elaborada pela autora.

Conforme Zheng (2014), a literatura indica uma proporção típica de itens de pré-teste no teste, entre 1/10 e 1/4. O autor também destaca que a decisão pode depender da necessidade de novos itens e de outros aspectos práticos.

O tamanho mínimo da amostra de respondentes também depende de vários aspectos como a exigência da precisão da estimativa, urgência na reposição de itens, etc (ZHENG, 2014). Também, deve-se considerar que, quando os testes são de alto impacto, não se deve expor desnecessariamente os itens. Por isso, é preciso ponderar entre precisão aceitável e mínima exposição possível dos itens aos respondentes.

Na literatura, não há um consenso sobre o tamanho da amostra mais conveniente, pois depende, dentre outros fatores, do modelo da TRI utilizado, dos itens, das próprias respostas dadas pelos respondentes e do método de estimação utilizado (MOREIRA JÚNIOR, 2011).

Um estudo feito por Nunes e Primi (2005) verificou o impacto do tamanho da amostra sobre a calibração de itens para os modelos de dois e três parâmetros da TRI. Os autores concluíram que, a partir de 200 respondentes é possível obter estimativas adequadas; já outros estudos trazem como aceitável um tamanho entre 500 e 1.000 respondentes para obter estimativas estáveis (GREEN et al., 1984; DEMARS, 2010; DE AYALA, 2009; WAINER; MISLEVY, 2000).

Stocking (1990) forneceu algumas orientações para a escolha da amostra de calibração para auxiliar na estimação dos parâmetros dos itens com maior precisão quando as propriedades não são conhecidas. Os níveis ótimos do traço latente para a calibração foram definidos com base na matriz de informação de Fisher para modelos ótimos (D-*optimal*).

Com base nesse estudo, Stocking (1990) concluiu que, se ML3P é considerado para análise de itens, então: (1) a amostra ideal e mais informativa para melhor estimar *a* é uma combinação de respondentes cujos traços latentes estão acima e abaixo do parâmetro *b* do item; (2) a amostra ideal e mais informativa para estimar o parâmetro *b* de itens fáceis e difíceis são os respondentes com traço próximo ao *b* do item; e (3) a amostra ideal e mais informativa para estimar *c* são os respondentes com baixo traço latente. Assim, para estimar todos os parâmetros dos itens, uma ampla distribuição dos níveis do traço latente (por ex., uniforme) é mais informativa do que uma distribuição em forma de sino.

Devido a esta problemática de que cada parâmetro exige uma amostra diferente para uma calibração mais precisa, a *optimal design theory* vem sendo muito utilizada, a qual lida com a definição de critérios de otimização para a estimativa simultânea de vários parâmetros. Segundo van der Linden e Ren (2015), a maior parte dos critérios

baseiam-se na matriz de covariância assintótica para os parâmetros que são estimados.

Um exemplo é a minimização do determinante da matriz de covariância ou, equivalente, a maximização do determinante da matriz de informação de Fisher; tal critério leva a soluções referidas como *D-optimal* [para mais detalhes, ver Berger (1994), van der Linden e Ren (2015), Zheng (2014), Guo (2016)]. Essa teoria auxilia a definir as amostras ideais em que, através de um plano de amostragem incompleta, diferentes amostras são atribuídas a diferentes itens, conforme ocorre no *design* adaptativo de calibração on-line (VAN DER LINDEN; REN, 2015; ZHENG, 2014).

- **Quais itens de pré-teste devem ser calibrados?**

Ao longo do tempo, nem toda edição de testes causará, obrigatoriamente, a necessidade de exclusão itens. É preciso acompanhar e detectar esta necessidade, identificando quais são as áreas prioritárias de reposição. Esta é uma tarefa de extrema importância e muito desafiadora. Por isso, a calibração de novos itens deve acontecer em toda edição de testes, seja para substituir itens superexpostos e/ou que apresentem alguns problema ou para melhorar a precisão da escala.

Nos casos de domínios (conteúdos) com pouca informação, é desejável expandir o BI focando nestes domínios para aumentar a quantidade de informação disponível em vários pontos da escala. Porém, conforme destacam Luecht, De Champlain e Nungester (1998), esta é uma tarefa não trivial, uma vez que requer que os itens sejam escritos para determinados níveis de dificuldade, assim como para satisfazer os requisitos do conteúdo.

As informações dadas pelo especialista no processo de elaboração dos itens (conteúdo e nível de dificuldade), podem auxiliar na tomada de decisão de quais itens novos devem ser selecionados para aplicação. Cabe destacar que, além desses requisitos, critérios psicométricos, técnicos e pedagógicos devem ser considerados para garantir clareza, qualidade na apresentação do item e nas alternativas de resposta, além de evitar erros de linguagem e gramática (ver BRASIL, 2010; HALADYNA, 2004).

Nesse contexto, a seleção dos itens de pré-teste dependerá muito do BI que cada programa de teste já tem à sua disposição. Análise da FIBI e uma análise descritiva da quantidade de itens nos diferentes níveis de dificuldade para os diferentes conteúdos avaliados podem auxiliar nesta decisão.

#### 3.1.1.1 *Design* de calibração on-line

A definição do *design* de calibração para itens de pré-teste envolve tomada de decisão sobre como serão obtidos os parâmetros dos itens. Os principais fatores para o *design* de calibração on-line são (ZHENG, 2014; GUO, 2016): **(I)** método de seleção de itens de pré-teste; **(II)** local de inserção do item de pré-teste no CAT; **(III)** método de estimação (algoritmo estatístico usado para estimar os parâmetros do item); e **(IV)** regra de parada para itens do pré-teste. Tais fatores nortearão as discussões a seguir, onde sugestões e estudos serão apresentados para auxiliar nas decisões referentes a esta etapa.

Uma das grandes vantagens do *design* de calibração on-line é que, como os itens novos são aplicados em conjunto com os itens operacionais no CAT, não é preciso utilizar um método de ligação; os itens utilizados para dar o traço latente ao respondente servem como ligação para os itens de pré-teste que estão sendo calibrados, ou seja, para colocar os itens na mesma escala de medida (VAN DER LINDEN; REN, 2015; ZHENG, 2014).

**I. Seleção de itens de pré-teste**

A seleção de itens de pré-teste pode ser na forma não-adaptativa (calibração não-adaptativa) ou adaptativa (calibração adaptativa).

**a) Calibração não-adaptativa**

Na seleção não-adaptativa de itens de pré-teste, nenhum tipo de adaptabilidade é utilizado no processo de calibração (ver KIM, 2006; BAN et al., 2001). Os itens de pré-teste são selecionados *a priori* pelo especialista, de um conjunto de itens disponíveis para serem pré-testados; quando o respondente atinge o local destinado a esses itens no CAT operacional (local de inserção), eles são aplicados até atingir a regra de parada. Após atingirem a regra de parada, um método de estimação é utilizado para calibrar os parâmetros dos itens.

Quando há muitos itens a serem calibrados e há um grande número de respondentes, é possível utilizar a técnica de Blocos Incompletos Balanceados (BIB), em que diferentes blocos de itens de pré-teste são utilizados para serem distribuídos entre os respondentes, possibilitando calibrar mais itens sem aumentar consideravelmente o comprimento do CAT (BEKMAN, 2001; NUNES et al., 2015). Ou então, pode-se aplicar de forma aleatória esses itens aos respondentes quando estes atingirem o

local de inserção, até que se obtenha um número razoável de respostas, capaz de estimar com precisão os parâmetros dos itens (ZHENG, 2014).

Quando o número de respondentes não for muito grande, deve-se aplicar os itens de pré-teste a todos os respondentes; caso o número de respondentes não seja suficiente, deve-se armazenar os dados e aplicá-los novamente na próxima edição do teste (STOCKING, 1988b).

**b) Calibração adaptativa**

Durante o CAT, o algoritmo seleciona adaptativamente, ou os itens de pré-teste com base na estimativa do traço latente do respondente, ou a amostra ao qual os itens de pré-teste serão apresentados (ALI; CHANG, 2014). Desta forma, diferentes itens de pré-teste podem ser selecionados para cada respondente, o que aumenta a segurança, pois reduz sua exposição (GUO, 2016, ZHENG, 2014).

Neste método de seleção, a calibração dos itens de pré-teste é executada junto com a aplicação do CAT. Particularmente, o *design* de calibração adaptativa objetiva aumentar a precisão das estimativas dos parâmetros dos itens com uma amostra menor do que quando utilizado métodos não-adaptativos (ALI; CHANG, 2014; KINGSBURY, 2009; CHANG; LU, 2010; GUO, 2016; ZHENG, 2014), principalmente para os itens que estão nos extremos da distribuição de dificuldade (KINGSBURY, 2009).

Estudos que abordam a calibração de itens de pré-teste no contexto de CAT e tentam, de alguma forma, fazer uso desta adaptabilidade para obter uma calibração mais eficiente são recentes. Alguns métodos de seleção de itens de pré-teste foram propostos para este contexto (GUO, 2016; ZHENG, 2014).

De forma geral, os estudos mostram-se promissores para a calibração de novos itens utilizando amostras pequenas. Citam-se van der Linden e Ren (2015); Kingsbury (2009), Zhu (2006), Ali e Chang (2014), Chang e Lu (2010), Makransky e Glas (2010), Lu (2014), Zheng (2014), Guo (2016).

A seleção adaptativa de itens de pré-teste pode ser centrada no respondente ou centrada no item, conforme classifica Zheng (2014):

- **Centrada no respondente:** Na seleção centrada no respondente, os itens de pré-teste são selecionados pelo mesmo método de seleção de itens usado para selecionar os itens operacionais no CAT (KINGSBURY, 2009). Todavia, Zheng (2014) destaca que esses métodos são projetados para otimizar a estimativa do traço latente do respondente, mas eles não são projetados para a finalidade de calibrar os itens do pré-teste. Logo, pode não ser

uma escolha razoável para o ML3P, pois as amostras ótimas para cada parâmetro do modelo variam. Por exemplo, os respondentes cujos níveis do traço latente correspondem ao nível de dificuldade do item geralmente fornecem menos informações para estimar os parâmetros *a* e *c* (ZHENG, 2014).

- **Centrada no item:** Análogo ao teste personalizado no CAT, onde um conjunto ótimo de itens operacionais é selecionado para cada respondente, a calibração on-line adaptativa centrada no item permite selecionar uma amostra ótima de respondentes para cada item de pré-teste, calibrando de forma mais eficiente os parâmetros do item (ZHENG, 2014).

Os métodos de seleção centrados nos itens combinam respondentes com itens de pré-teste a partir de critérios diretamente projetados para otimizar a estimativa dos parâmetros do item de pré-teste (ZHENG, 2014). Um dos critérios frequentemente adotados é o critério de D-*optimal* (BERGER, 1994; CHANG; LU, 2010; GUO, 2016; VAN DER LINDEN; REN, 2015; ZHENG, 2014).

No entanto, Zheng (2014) ressalta que a maioria dos estudos adotam o *design* ótimo tradicional, o que é praticamente impossível na calibração adaptativa, pois é assumido que há um "banco de respondentes" composto por indivíduos em diferentes níveis do traço latente e, para cada item de pré-teste, **os respondentes são comparados** e aqueles cujos níveis do traço latente maximizam o critério de D-*optimal* (isto é, fornece a maior informação para o item) são selecionados.

Essa prática requer que todos os respondentes completem os seus testes operacionais na Fase 1 e formem um "banco de respondentes" a partir do qual os indivíduos ótimos são selecionados durante a Fase 2. Porém, isso não ocorre em CATs, visto que os testes operacionais são administrados em momentos dispersos dentro de uma janela de teste, não havendo um "banco de respondentes" estático para escolher o respondente e, também, é desconhecido o nível do traço latente que o próximo respondente está (ZHENG, 2014). Exemplo disso é o *design* em dois estágios de Chang e Lu (2010).

Van der Linden e Ren (2015) implementaram um procedimento mais viável, propondo que **todos os itens do pré-teste sejam comparados** e o item com o valor da estatística D-*optimal* bayesiana seja selecionado quando o respondente atigir o local de inserção do item no CAT operacional. Esta estatística baseia-se na contribuição *a posteriori* esperada dos respondentes à matriz de informação observada para os parâmetros dos itens.

Para Zheng (2014), o método desenvolvido por Van der Linden e Ren (2015) também tem sua limitação, uma vez que alguns itens de pré-teste tendem a produzir valores D-*optimal* maiores do que outros devido à sua superioridade estatística, fazendo com que esses itens sejam mais selecionados, mesmo que outros itens precisem mais do respondente atual para a calibração. O impacto disso é que, ao finalizar a fase de calibração, o desenvolvedor poderá obter alguns itens com estimativas precisas, mas outros com estimativas pouco confiáveis ou sem estimativas de parâmetros do item (ZHENG, 2014).

Outro método existente de seleção centrada em itens é o índice de adequação (SI - *suitability index*) proposto por Ali e Chang (2014). Segundo os autores, esse índice é baseado nos tamanhos de amostra alvo de uma determinada gama de traço latente para o item e nos tamanhos de amostra atuais que já responderam o item de pré-teste, considerando vários intervalos de traço latente particionados. Quando um examinado atinge os locais de inserção, o item do pré-teste que maximiza o índice SI será selecionado. Este *design* procura controlar o tamanho da amostra a partir de diferentes intervalos do traço latente para cada item de pré-teste; em estudos simulados, os resultados mostram-se promissores, mas a construção destes índices é arbitrária (ZHENG, 2014).

Zheng (2014) propôs um novo método de seleção de itens de pré-teste, denominado Índice de Prioridade de Intervalo Informativo Ordenado (OIRPI - *Ordered Informative Range Priority Index*) com dois algoritmos para implementá-lo: OIRPI com estatísticas de ordem e OIRPI com padronização. O objetivo do método OIRPI é atribuir o respondente atual ao item de pré-teste que mais precisa dele; o método determina o quanto um item precisa desse respondente "por quão informativo este respondente é a este item, em comparação com os respondentes, em outros níveis do traço latente" (ZHENG, 2014).

A seguir, apresentam-se alguns algoritmos desenvolvidos para este método de calibração adaptativa de itens de pré-teste.

**Algoritmos para calibração adaptativa**

Até o momento, não encontrou-se *softwares* disponíveis para fins comerciais que fazem uso da calibração adaptativa. Neste caso, é preciso implementar o algoritmo. Na literatura, são propostos alguns deles, os quais serão citados a seguir.

Etapas gerais do *design* de calibração adaptativa proposto por Zheng (2014) é dado por:

**1.** Inicialize os parâmetros do item de pré-teste:

- Opção 1: A amostragem aleatória pode ser usada para cada item de pré-teste até que um tamanho de amostra de calibração mínimo predeterminado seja atingido. Em seguida, os itens de pré-teste são calibrados usando esses dados de amostragem aleatória e esses parâmetros são usados como os parâmetros iniciais do item para a seleção adaptativa na Etapa 2.
- Opção 2: Cada item de pré-teste pode ser classificado em vários níveis de dificuldade por especialistas em conteúdo. Os parâmetros de dificuldade podem então ser inicializados com base nesta classificação e os parâmetros de discriminação ou acerto ao acaso, se o modelo exigir, podem ser inicializados com os valores mais comuns (Ex., $a = 1$, $c = 0{,}2$ para cinco alternativas de resposta);

**2.** Durante o CAT operacional, quando um respondente atingir os locais de inserção do item de pré-teste, o algoritmo selecionará e administrará o item de pré-teste mais desejável do conjunto de itens disponível, conforme regra de seleção de item escolhida. Os locais de inserção podem ser predeterminados e fixados ou escolhidos aleatoriamente dentro de um certo intervalo;

**3.** Quando um respondente completar o seu CAT, o algoritmo usará um dos métodos de estimação estatística predefinido para atualizar os parâmetros do item de pré-teste administrado. Alternativamente, quando um item de pré-teste tiver obtido uma quantidade suficiente de novos dados de resposta (ex., 10 respostas), seus parâmetros do item serão atualizados. Para cada item a ser estimado, todos os dados de resposta são usados para o procedimento de estimação;

**4.** Passos 2 e 3 são executados para cada indivíduo. Os itens do pré-teste podem ser exportados individualmente da fase de pré-teste, assim que a regra de parada para o item for atingida. A regra de parada pode ser baseada em tamanhos de amostra ou precisão de medição. Em seguida, a iteração dos Passos 2 e 3 continua com os itens restantes do pré-teste (ou o conjunto de itens do pré-teste pode ser reabastecido com outros itens de pré-teste) até que a edição de teste seja encerrada.

Ali e Chang (2014) propuseram o *design* denominado *Item-driven adaptive*. O algoritmo utiliza blocos de itens ou itens individuais, conforme os seguintes passos:

**1.** Comece a administrar esses itens do pré-teste a uma amostra de respondentes de tamanho $N$ (pode ser de forma aleatória);

2. Obtenha uma estimativa inicial dos parâmetros do item por meio da calibração pela TRI (usou ML3P);
3. Agrupe esses itens em uma série de blocos (por exemplo, cinco blocos), que diferem na dificuldade média, utilizando os resultados do Passo 2;
4. Para cada respondente, escolher de forma adaptativa um bloco ou itens individuais, conforme a estimativa do traço latente ou pela estimativa da dificuldade, ou por meio de um outro critério adequado para selecionar itens do pré-teste, e administrar o bloco;
5. Atualize as estimativas dos parâmetros do item com base em dados obtidos a partir do Passo 4;
6. Repita os passos 4 e 5 até que estimativas estáveis dos parâmetros do item sejam obtidas (por ex.: mínima mudança na estimativa dos parâmetros dos itens), ou satisfaça uma outra regra de parada (por ex.: tamanho da amostra).

No método proposto por Ali e Chang (2014), o *design* é flexível na medida em que ele pode ser completamente adaptativo, onde itens individuais são selecionados e aplicados ou, alternativamente, ter vários blocos, os quais podem ser formados considerando o conteúdo dos itens e impossibilitando que o respondente receba apenas os itens de um mesmo conteúdo, ou então, cada bloco pode compreender vários itens adequados para um determinado intervalo do traço latente, o que corresponde, de certo modo, aos testes multi-estágios.

Kingsbury (2009) apresentou o algoritmo para o modelo de um parâmetro da TRI:

1. Estabeleça um conjunto de itens de pré-teste a serem aplicados. Esses itens ainda não têm respostas;
2. Forneça uma estimativa inicial provisória da dificuldade para cada item (isto pode ser feito por especialistas em conteúdo);
3. Defina as regras para posicionamento/inserção desses itens dentro do CAT;
4. Quando as regras indicarem que deve ser aplicado um item de pré-teste ao respondente, o item é escolhido a partir do conjunto como sendo aquele que fornece mais informação com base na dificuldade provisória do item e na estimativa momentânea do traço latente do respondente. Restrições adicionais podem ser aplicadas aqui. Por exemplo, de conteúdo;

5. Depois de um particular item de pré-teste ter sido aplicado a um número pré-especificado de indivíduos, a estimativa provisória de dificuldade do item é atualizada por meio do método de estimação desejado. Esta nova estimativa provisória da dificuldade é então usada para calcular as informações do item, que será utilizada como informação para uma posterior seleção dos itens e aplicação;
6. Passos 4 e 5 são repetidos até que um número predefinido de respostas aos itens seja obtido ou até que a estimativa provisória da dificuldade estabilize a um nível predeterminado;
7. Para o item em questão, a estimativa provisória final da dificuldade do item torna-se a estimativa da dificuldade do item operacional (desde que o item atenda a todos os requisitos estatísticos);
8. O processo continua até que todos os itens sejam calibrados.

**Resultados de alguns estudos que utilizaram a calibração adaptativa**

Os resultados de Kingsbury (2009) e Chang e Lu (2010) indicaram que é possível avaliar as características do item com menos respondentes do que na apresentação aleatória e que os ganhos são particularmente visíveis para os itens nos extremos da distribuição de dificuldade.

Makransky e Glas (2010) desenvolveram estratégias para, simultaneamente, calibrar itens e estimar o traço latente dos respondentes, sem usar informações prévias. No entanto, o estudo apresenta várias limitações sobre os métodos e os resultados são direcionados à avaliar a precisão das estimativas do traço latente.

Ali e Chang (2014) utilizaram 12 itens operacionais e 15 itens de pré-teste disponíveis para aplicação. Destes, somente três itens eram aplicados a cada respondente, totalizando um comprimento do teste de 15 itens. Esses três itens novos foram inseridos próximo ao final do teste e três métodos de seleção de itens foram comparados (*maximum* SI; diferença mínima entre a estimativa atual do traço latente do respondente e a dificuldade do item; e seleção aleatória).

Os autores utilizaram método de máxima verossimilhança marginal com vários ciclos EM (MEM) para calibração dos itens de pré-teste, atualizando as estimativas após atingir tamanhos de amostras especificadas. Os resultados deste estudo forneceram evidências de que um *design* adaptativo orientado para o item em pré-teste gera menor viés e estimação mais precisa, mesmo para amostras pequenas. Além disso, o método SI apresentou melhor desempenho por produzir uma amostra mais adequada.

Zheng (2014) utilizou o tamanho da amostra de 500 respostas como regra de parada para cada item do pré-teste e as estimativas dos parâmetros dos itens eram atualizadas a cada 10 respostas obtidas. De forma geral, os dois métodos OIRPI propostos apresentaram melhor eficiência em comparação aos métodos D-*optimal*, seleção centrada no respondente e seleção aleatória.

Nos estudos de Van der Linden e Ren (2015), os resultados de um estudo simulado indicaram viabilidade do algoritmo *D-optimal* bayesiano para a calibração de itens. Uma comparação das performances *D-optimal* com outros critérios ótimos e atribuição aleatória de itens mostrou uma calibração mais rápida de um número substancial de itens para o critério da *D-optimal* bayesiano.

## II. Local de inserção dos itens de pré-teste no CAT

Uma vez definido quais itens de pré-teste serão aplicados, é preciso definir as regras de inserção no CAT (local), as quais devem ser implementadas no algoritmo computacional, tanto para o *design* de calibração adaptativa quanto não-adaptativa. Várias regras de inserção de itens podem ser encontradas na literatura.

Masters, Muckle, Bontempo (2009) sugerem que itens novos sejam aplicados aleatoriamente a respondentes entre itens operacionais de um CAT. Bock, Muraki e Pfeiffenberger (1988) sugerem que os itens sejam acrescentados ao final dos itens operacionais para não influenciar no traço latente. Stocking (1994) sugere acrescentar os itens no meio ou no final do teste. Em Segall (2003), a posição dos itens do pré-teste é escolhida aleatoriamente para cada respondente como o segundo, terceiro ou quarto item do CAT; cada respondente recebe um item de pré-teste ao longo de 10 itens operacionais, por exemplo.

No contexto de calibração adaptativa, Ali e Chang (2014), van der Linden e Ren (2015) e Zheng (2014) sugerem que os itens sejam acrescentados ao final do teste, pois a informação sobre o traço latente do respondente é mais precisa e esta informação é utilizada para selecionar os itens de pré-teste. Porém, van der Linden e Ren (2015), Davey, Pommerich e Thompson (1999) e Zheng (2014) ressaltam que o uso permanente desta regra pode tornar os itens novos reconhecíveis e, consequentemente, aumentar a probabilidade de um comportamento "menos sério" do respondente no final do teste, gerando um *drift* entre o pré-teste e o teste operacional.

Uma solução prática dada por van der Linden e Ren (2015) é atribuir alguns itens em posições próximas ao final do teste. Kingsbury

(2009) também sugere uma regra de que "não mais de um item de pré-teste seja administrado consecutivamente". Zheng (2014) comparou as posições desses itens inseridos aleatoriamente no início, no meio e no final do CAT para a calibração adaptativa, uma vez que as estimativas do traço latente nesses estágios do teste possuem diferentes níveis de precisão. Os resultados indicaram que a inserção dos itens de pré-teste no meio e final do teste conduzem a resultados de calibração mais precisos.

Davey, Pommerich e Thompson (1999) discutem a importância da regra de parada do CAT operacional na definição do local de inserção dos itens de pré-teste. Os autores sugerem que predizer o comprimento do CAT (via simulação) pode ajudar a decidir sobre como será a inserção desses itens de pré-teste no CAT, principalmente para CAT de comprimento variável. Além disso, eles enfatizam que deve-se ter cuidado quando se estabelece tempo de teste em CAT, uma vez que em testes acelerados, a não resposta aos itens pode prejudicar tanto o traço latente do respondente quanto a calibração de itens de pré-teste.

## III. Métodos de estimação dos parâmetros dos itens de pré-teste

Muitos estudos concentraram-se sobre os métodos de estimação, principalmente sobre o método de máxima verossimilhança marginal (MML) com algoritmo EM (BOCK; AITKIN, 1981; DEMPSTER; LAIRD; RUBIN, 1977; KIM, 2006). A ideia é utilizar os valores de parâmetros conhecidos dos itens operacionais para auxiliar na estimação dos itens de pré-teste, que conforme Zheng (2014), o problema de estimativa na calibração on-line é essencialmente o mesmo que na calibração de parâmetros fixos, na qual os parâmetros do item operacional são fixados e os parâmetros do item do pré-teste são calibrados.

Métodos utilizados na calibração on-line são: MML com vários ciclos EM (denominado MEM) (BAN et al., 2001); MML com um ciclo EM (denominado OEM) (WAINER; MISLEVY, 2000); Método-A e Método-B (STOCKING, 1988b), método MCMC (SEGALL, 2003), métodos MML e não-paramétricos (KRASS; WILLIAMS, 2003), adaptações bayesianas dos métodos Método-A, OEM e MEM (ZHENG, 2014) e BILOG com priori informativa (BAN et al., 2001). Zheng (2014) apresentou um resumo dos métodos, conforme segue:

- Método Stocking-A (STOCKING, 1988b): calcula o traço latente dos respondentes usando todos os itens operacionais administrados e, em seguida, estima os parâmetros do item de pré-teste usando valores da estimativa de máxima verossimilhança condicional. Stocking-A é o método mais

simples, mas pode resultar em *drift* da escala por causa da utilização do traço latente estimado do respondente em vez dos valores reais (STOCKING, 1988b).

- Método Stocking-B (STOCKING, 1988b): é semelhante ao Stocking-A, mas com um passo de equalização usando itens âncora para corrigir o *drift* da escala. Teoricamente é mais rigoroso e mais difícil do que Stocking-A, pois a necessidade de itens âncora aumenta o comprimento do teste e o tamanho da amostra de calibração (BAN et al., 2001).
- Método OEM (WAINER; MISLEVY, 2000): obtém-se primeiro a distribuição *a posteriori* do traço latente utilizando todos os itens operacionais administrados. Essa distribuição é usada para marginalizar a função de verossimilhança e estimar os parâmetros do item de pré-teste. Os parâmetros do item estimados são aqueles que maximizam a função de verossimilhança esperada *a posteriori*.
- Método MEM (BAN et al., 2001): o primeiro ciclo é o mesmo que OEM. No segundo ciclo, os itens operacionais e de pré-teste são usados para atualizar a distribuição *a posteriori* do traço latente para a estimação OEM. A iteração EM continua à medida que os parâmetros do item de pré-teste são atualizados. Considera-se que a iteração convergiu se as estimativas diferem por não mais que um pequeno limiar.
- Método BILOG com priori informativa (BAN et al., 2001): utiliza o *software* BILOG (ZIMOWSKI et al., 2003) para calibrar itens de pré-teste em uma única execução. Ele fixa os parâmetros dos itens operacionais, definindo distribuições *a priori* bastante informativas para esses itens. Em seguida, ele calibra os itens de pré-teste e operacionais simultaneamente.

O estudo de Ban et al. (2001) comparou vários métodos por meio de simulações, na calibração não-adaptativa. Os resultados mostraram que o método MEM foi o melhor porque produziu os menores erros de estimativa dos parâmetros para todas as condições de tamanho de amostra avaliados. MEM também foi preferível a OEM, a não ser que o período de tempo envolvido no cálculo iterativo seja uma preocupação. O Método-B também funcionou muito bem, mas são necessários itens comuns de ligação, o que aumentaria o comprimento do teste ou requer amostras maiores, dependendo do *design* de aplicação do teste.

Os autores também concluem que o método Bilog pode não ser uma escolha razoável para pequenas amostras e o Método-A apresentou

o pior desempenho. Todos os itens do pré-teste foram aplicados ao mesmo grupo de respondentes. Stocking (1988b) também mostrou que o Método-B é superior ao Método-A em termos de propriedades estatísticas.

Na calibração adaptativa, Zheng (2014) comparou seis métodos de estimação: Stocking-A, OEM, MEM e suas versões bayesianas, como *Bayesian* Stocking-A, *Bayesian* OEM e *Bayesian* MEM. O autor recomenda o uso do método de estimação *Bayesian* MEM.

Ban et al. (2002) comparam os métodos OEM, MEM e Método-B em termos de recuperação dos parâmetros dos itens de pré-teste quando as respostas são escassas tanto para os itens do pré-teste (por ex., se usar BIB) quanto para os itens operacionais do CAT. Simulações do CAT foram utilizadas para avaliar os resultados e o método MEM produziu o menor erro médio total da estimativa dos parâmetros; o método OEM produziu os piores resultados.

Segall (2003) avaliou procedimentos de estimação alternativos, calibrando simultaneamente tanto itens de pré-teste quanto itens operacionais. O autor afirma que reestimar todos os parâmetros ao mesmo tempo, pode diminuir o problema de *drift* da escala. Para fixar a escala de medida, presume-se que os parâmetros de pelo menos alguns itens permanecem constantes ao longo do tempo e são conhecidos. Os resultados de um estudo de simulação indicaram que o procedimento pode proporcionar boa recuperação das FRIs e estimar suficientemente bem os parâmetros dos itens de pré-teste para dados faltantes, resultantes do CAT.

Por outro lado, alguns autores afirmam que a calibração simultânea em conjunto com um método de ligação pode levar a vários problemas devido à natureza adaptativa de obtenção de dados no CAT (GONZÁLEZ-BETANZOS; ABAD; BARRADA, 2014; KIM, 2006; BAN et al., 2001). Por isso, Kim (2006) propôs o método FIPC (*fixed item parameter calibration*), no qual após a obtenção de um número suficiente de respondentes, itens operacionais têm seus valores fixados e conhecidos, sendo utilizados como elementos âncora, permitindo que os parâmetros dos itens de pré-teste estejam na mesma escala métrica dos itens operacionais (GONZÁLEZ-BETANZOS; ABAD; BARRADA, 2014; KIM, 2006). Este método mostrou-se bem sucedido na calibração dos itens do pré-teste no ambiente CAT (BAN et al., 2001).

Um ponto importante a ser destacado é que, uma vez definido o método de estimação, este deve ser utilizado ao longo do tempo pelas organizações de testes.

**IV. Regra de parada na calibração on-line**

Esta regra determina quando parar a amostragem para os itens de pré-teste. Desta forma, são semelhantes às regras utilizadas para encerrar o CAT operacional. A regra de parada mais simples é baseada em tamanhos de amostra predeterminado (ALI; CHANG, 2014; KINGSBURY, 2009; ZHU, 2006). No entanto, Zheng (2014) destaca que, com o mesmo tamanho de amostra, parâmetros de itens diferentes podem ter erros padrão diferentes. Sendo assim, regras de parada podem ser definidas com base na precisão, a qual deverá ser mais eficiente do que a regra com base no tamanho da amostra.

Outra possível regra é combinar o EP com uma amostra de tamanho máximo para impedir a "infinita duração" da aplicação dos itens. Isso é especialmente útil na calibração adaptativa. Kingsbury (2009) propôs parar a amostragem quando as estimativas dos parâmetros do item se estabilizarem.

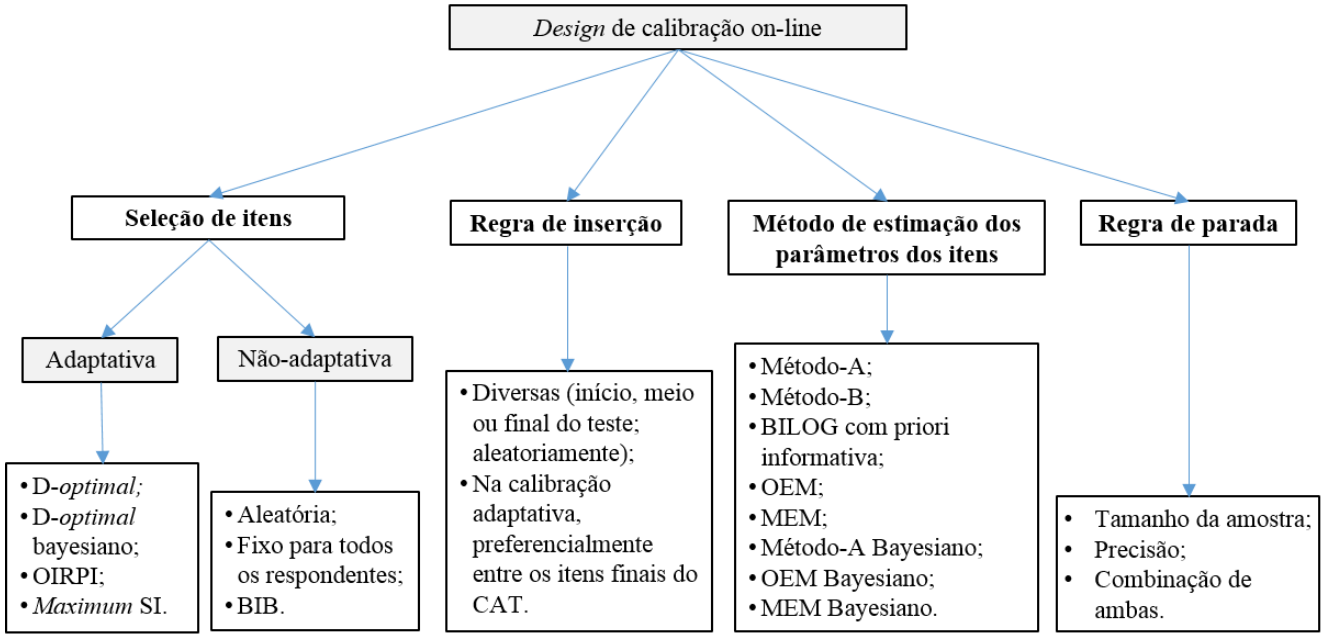
Para finalizar a Fase 1 da sistemática, a Figura 8 mostra um resumo dos métodos abordados para operacionalizar a calibração on-line de itens de pré-teste.

3.1.1.2 Critérios para avaliar a qualidade dos itens pré-testados

Nesta etapa, apresentam-se os critérios que podem auxiliar na tomada de decisão de incluir os itens pré-testados no BI ou quando eles não devem ir para o BI, ou seja, devem ser excluídos (ou reelaborados pelos especialistas), após a obtenção das estimativas. Assim como no desenvolvimento do BI inicial para CAT, os novos itens precisam atender aos pressupostos da TRI, apresentar parâmetros dos itens satisfatórios e precisos, e não apresentar DIF significante.

Um bom item deve ser confiável e válido, de modo a contribuir idealmente para a confiabilidade interna do teste e discriminar bem os respondentes. Nesta etapa, a CCI e a FII auxiliam nas decisões de quais itens estão funcionando bem e quais apresentam problemas (COHEN; SWERDLIK; STURMAN, 2014).

Figura 8 – Principais fatores e métodos para o *design* de calibração on-line de itens de pré-teste.




Fonte: Elaborada pela autora.

**I. Precisão e bons parâmetros**

A precisão das estimativas é um importante aspecto a ser considerado na calibração dos itens, que dependerá muito dos itens em si (por ex., escrita e dificuldade) e da amostra de respondentes. Portanto, deve-se considerar que:

- EP elevado pode estar relacionado ao tamanho da amostra insuficiente para obter estimativas estáveis ou das características da amostra utilizada para calibração. Neste caso, observar a porcentagem de respostas corretas pode ser importante, bem como obter uma amostra maior e mais diversificada de respondentes, se for o caso;
- Se a amostra for grande e o EP elevado, pode indicar má formulação do item, tanto no enunciado quanto nas alternativas de respostas. Neste caso, é fundamental o auxílio de especialistas de conteúdo;
- O EP é influenciado pela magnitude dos valores dos parâmetros, por isso, quando este valor for elevado, deve-se comparar com valores semelhantes de outros itens, a fim de diagnosticar alguma incoerência.

Em relação aos valores dos parâmetros de cada item na escala N(0,1), é esperado que:

- Os valores de *a* devem ser positivos, indicando que a probabilidade de resposta correta aumenta quando o traço latente aumenta (ANDRADE; TAVARES; VALLE, 2000), sendo adequado valores acima de 0,7 (TEZZA et al., 2011). Valor muito baixo (próximo a zero) indica que a probabilidade de acertar o item é semelhante para quem tem alto ou baixo nível do traço latente; também, pode ser um indicativo de que o item pertence a uma dimensão secundária do traço latente (GREEN et al., 1984; TEZZA et al., 2011). Em CAT, itens com altos valores de *a* tornam o CAT mais eficiente porque menos itens serão necessários para cada respondente (GREEN et al., 1984);
- Valores de *b* comumente variam de -4 a 4 na escala N(0,1);
- Para um item com cinco alternativas de resposta é esperado valores em torno de 0,2. Valores muito discrepantes, por exemplo, $c = 0,5$, devem ser investigados por especialistas de conteúdo.

Por fim, deve-se ficar atento ao EP das estimativas, bem como aos valores dos parâmetros estimados no pré-teste. Havendo incoerências, os especialistas de conteúdo devem ser consultados e o item não deve ir para o BI. Também, a investigação dos pressupostos da TRI e do DIF podem dar mais subsídios para essa decisão.

## II. Pressupostos da TRI em CAT

Poucos trabalhos tem sido direcionados especificamente a esta área do CAT. Atualmente, técnicas utilizadas em testes convencionais, também são utilizadas nos testes adaptativos, conforme apresentado na seção 2.2.2.1. Quando é diagnosticado o não atendimento dos pressupostos, alguns possíveis "caminhos" para lidar com isso são propostos na literatura.

Para verificar a unidimensionalidade do BI, comumente considera-se que, se o modelo da TRI se ajusta adequadamente aos dados de resposta e se fatores secundários (análise fatorial de informação completa) não exibem nenhum padrão discernível, a unidimensionalidade pode ser assumida (GREEN et al., 1984).

É possível utilizar modelos multidimensionais da TRI, mas advertências devem ser observadas. De acordo com Thissen et al. (2007), a complexidade da modelagem multidimensional aumenta em comparação a TRI unidimensional e a interpretação dos traços latentes é mais complexa. Para Piton Gonçalves e Aluísio (2015), o desenvolvimento e a manutenção de um CAT multidimensional são mais caros e trabalhosos que em CAT unidimensional e o tempo computacional de processamento poderá ser muito longo quando há um BI grande para medir as diferentes dimensões.

Uma alternativa para lidar com a multidimensionalidade é excluir itens de modo a tornar o teste unidimensional ou, então, separar os itens em vários bancos unidimensionais (conforme os conteúdos) e, em seguida, pode ser aplicado subtestes ou CATs curtos separados (chamados de mini-CAT) para cada BI unidimensionais (GREEN et al., 1984; THOMAS, 1990; SCHNIPKE; GREEN, 1995).

Neste caso, o teste final é criado pela combinação de um número de mini-CATs (SCHNIPKE; GREEN, 1995) e a montagem de multi-escalas não impede o cálculo posterior de um único traço latente (por ex., pela média), se isso for útil como um resumo ou índice do resultado do teste (THISSEN et al., 2007; THOMAS, 1990).

Kalinowski, Natesan e Henson (2014) investigaram a penalidade de utilizar CAT unidimensional quando dados são multidimensionais. Os resultados indicaram estimativas de parâmetros dos itens com

relativamente grandes erros padrão. Os autores recomendam utilizar CAT unidimensionais somente se recalibração como modelo multidimensional for muito caro.

Tipos de itens que violam a propriedade do item de independência local devem ser evitados (GREEN et al.,1984). Porém, restrições podem ser impostas no algoritmo de modo que itens que se sobrepõem ou que são dependentes não aparecem no mesmo teste; eles são considerados como itens inimigos e, desta forma, podem ser aproveitados (WAY, 2006; ALI; CHANG, 2014). Alguns autores sugerem o uso de *testlet* em CAT (MURPHY; DODD; VAUGHN, 2010; DE AYALA; 2009; DEMARS, 2010).

*Testlet* é definido como um grupo de itens relacionados a uma única área de conteúdo ou estímulo, onde testes são construídos selecionando *testlets* ao invés de itens separadamente (SCHNIPKE; GREEN, 1995). Nesses casos, modelos para *testlet* e métodos adequados podem ser utilizados para estimação dos parâmetros, uma vez que as respostas ao item dentro de um *testlet* não são localmente independente (MURPHY; DODD; VAUGHN, 2010; SIRECI; WAINER; THISSEN, 1991).

### III. Métodos para detectar DIF em CAT

Na seção 2.2.1.1, citaram-se os métodos utilizados para análise do DIF em contextos de testes tradicionais. No entanto, esses critérios comumente utilizados para análises DIF, baseados no traço latente total (*number-correct score*), tornam-se sem sentido pela natureza adaptativa dos CATs (LEI; CHEN; YU, 2006; NANDAKUMAR; ROUSSOS, 2004). Isso porque em um CAT ideal, ou seja, com itens ilimitados disponíveis em todos os níveis do traço latente, é esperado obter o mesmo número de respostas corretas de cada respondente, independente do nível onde ele se encontra na escala (NANDAKUMAR; ROUSSOS, 2004).

Também, a dificuldade de análise do DIF no ambiente CAT acontece devido à matriz incompleta de respostas para os itens operacionais, pois diferentes respondentes podem responder a itens diferentes e itens diferentes podem receber um número diferente de respostas, em conjunto com uma gama restrita do nível do traço latente dos respondentes em cada item (GONZÁLEZ-BETANZOS; ABAD; BARRADA, 2014; LEI; CHEN; YU, 2006). Logo, se distribuições dos grupos focal e referência para o traço latente diferem, deve haver pouca sobreposição itens operacionais (GONZÁLEZ-BETANZOS; ABAD; BARRADA, 2014).

Detectar itens com DIF é importante porque eles podem invalidar os procedimentos para a tomada de decisões sobre os indivíduos (GONZÁLEZ-BETANZOS; ABAD; BARRADA, 2014). Alguns autores (ZWICK; THAYER; WINGERSKY, 1993, 1994a; ZWICK, 2010; MAKRANSKY; GLAS, 2013) afirmam que a detecção de DIF é mais importante em CAT do que na forma tradicional, sobretudo pelo modo de aplicação do teste (via computador), que cria fontes potenciais de DIF que não estão presentes em testes P&P, como as diferentes familiaridades com o computador por parte dos respondentes. Além disso, poucos itens são aplicados ao respondente em CAT e respondentes são comparados com base em suas respostas aos diferentes itens, desempenhando um papel mais importante.

Alguns estudos foram desenvolvidos visando atender a essa demanda de análise de DIF em CAT (Quadro 11), os quais apresentam uma adaptação nos testes já utilizados, como MH, *standardization*, SIBTEST e RL, para usar a estimativa do traço latente do CAT (LEI; CHEN; YU, 2006; NANDAKUMAR; ROUSSOS, 2004) ou o traço latente verdadeiro esperado ao longo de todo o BI baseado na estimativa do traço latente CAT, obtida com a aplicação de alguns itens operacionais (ZWICK, THAYER, WINGERSKY, 1993, 1994a, 1994b).

Poucos estudos analisam o DIF em itens de pré-teste em CAT, como González-Betanzos, Abad e Barrada (2014), Lei, Chen e Yu (2006); Nandakumar e Roussos (2004) e Zwick, Thayer e Wingersky (1993, 1994b). Nesses estudos, os itens de pré-teste foram aplicados a todos os respondentes para uma calibração não-adaptativa e diferentes variáveis são manipuladas.

Em um estudo feito por Zwick, Thayer e Wingersky (1994a), os autores concluíram que o tamanho das estimativas do DIF eram afetadas, em geral, pelo tamanho do verdadeiro DIF, pela dificuldade do item e as interações do verdadeiro DIF com a dificuldade e a discriminação do item; quase nenhum efeito foi encontrado para a distribuição do grupo focal, para a posição item e tamanho da amostra.

González-Betanzos, Abad e Barrada (2014) utilizaram a técnica de imputação de dados para itens não respondidos em CAT; porém, os autores concluíram que esta abordagem não é necessária (e deve ser evitada), se os parâmetros dos itens operacionais são fixos e os pressupostos da TRI são assegurados, visto que poderá introduzir viés nos parâmetros dos itens e na distribuição do grupo na escala, quando o traço latente CAT for estimado com níveis elevados de erro.

Quadro 11 – Estudos sobre DIF para itens operacionais e de pré-teste em CAT.

| Autor | Método DIF | Objetivo | Variáveis manipuladas | Resultados |
|---|---|---|---|---|
| González-Betanzos, Abad e Barrada (2014)<br><br>**DIF para itens de pré-teste em CAT** | IRT-LRT | Por meio de simulação, analisar o desempenho do método IRT-LRT na detecção de DIF em itens de pré-teste aplicados junto ao CAT, usando o método de calibração FIPC (KIM, 2006), com imputação (LEI; CHEN; YU, 2006) e sem imputação de respostas para itens não respondidos. | Tipo de DIF (sem DIF, DIF uniforme e não-uniforme), tamanho do DIF (moderado, grande), tamanho do impacto (distribuição do traço latente), comprimento do teste e tamanho da amostra. | Os resultados com e sem imputação foram analisados em relação à detecção de DIF, recuperação dos parâmetros de distribuição de cada grupo e do tamanho do DIF. O método FIPC com múltiplas atualização e múltiplos ciclos EM mostrou-se adequado para detecção de DIF de grande porte, em grandes amostras. Na presença de impacto, o uso de imputação levou a um viés no tamanho do DIF, não sendo recomendável sua utilização em CAT de comprimento curto. |
| Nandakumar e Roussos (2004)<br><br>**DIF para itens de pré-teste em CAT** | CATSIB | Propor e avaliar por meio de simulação o método CATSIB, que é uma versão modificada do SIBTEST (método usado para detectar DIF em P&P) com correção da regressão, para evitar uma elevada taxa de erro Tipo I. | Tamanho da amostra, tamanho do impacto e tamanho do DIF. | É um método prático e confiável para a detecção de DIF em itens de pré-teste que são calibrados em um ambiente CAT e pode ser generalizado para itens operacionais. CATSIB com correção da regressão fornece bom controle do erro Tipo I e o poder de detecção de DIF é bem maior para amostras grandes (500) do que para amostras menores (250). Enquanto CATSIB, sem correção da regressão, apresentou erro Tipo I inflacionado. |

Continuação

| Autor | Método DIF | Objetivo | Variáveis manipuladas | Resultados |
|---|---|---|---|---|
| Lei, Chen e Yu (2006)<br><br>**DIF para itens de pré-teste em CAT** | RL e IRT-LRT adaptados para CAT; e CATSIB com correção da regressão. | Propor e analisar, por meio de simulação, o desempenho de IRT-LRT e RL adaptados para CAT, comparando-os com CATSIB. Imputação de dados para os itens não respondidos foi utilizado. | Tipo de DIF, tamanho do impacto e tamanho da amostra. | Os métodos modificados foram mais poderosos do que CATSIB para detecção de DIF não-uniforme. CATSIB foi melhor para detectar DIF uniforme. IRT-LRT mostrou baixo poder de detecção em itens muito difíceis e pouco discriminativos com DIF não-uniforme. IRT-LRT forneceu adequado controle do erro Tipo I nas condições testadas, já os outros métodos foram afetados pelo tamanho da amostra e impacto. No entanto, ele não é um método muito prático. |
| Zwick, Thayer e Wingersky (1993)<br><br>**DIF para itens de pré-teste e operacionais em CAT** | Versões MH e *standardization* adaptados para CAT. | Investigar o desempenho das versões modificadas dos métodos MH e *standardization* em CAT, e comparar com a versão padrão (sem modificações) desses métodos (DORANS; KULICK, 1986; HOLLAND; THAYER, 1988), para itens operacionais e de pré-teste. | Tipo de DIF (sem DIF, DIF uniforme – não correlacionado e correlacionado positivamente com a dificuldade do item), tamanho da amostra e tamanho do impacto. | Alta correlação foi obtida entre as versões adaptadas para CAT operacional e DIF verdadeiro ou para as versões padrão. Para os itens de pré-teste, obteve-se alta correlação com o verdadeiro DIF, mas com magnitude um pouco menor do que as estatísticas para CAT operacional. Além disso, os erros padrão de MH para itens de pré-teste foram maiores do que para o CAT, reduzindo o poder de detecção. |

Continuação

| Autor | Método DIF | Objetivo | Variáveis manipuladas | Resultados |
|---|---|---|---|---|
| Zwick, Thayer e Wingersky (1994b)<br><br>**DIF para itens de pré-teste em CAT** | Versões MH e *standardization* adaptados para CAT. | É uma extensão do estudo de Zwick, Thayer e Wingersky (1993), que visa avaliar a utilização de métodos alternativos de correspondência para os itens do pré-teste, ao invés de usar a soma dos traços latentes verdadeiros esperados com base nas respostas ao CAT e o traço latente (0 ou 1) para o item em análise (como no estudo anterior). | Tipo de DIF (sem DIF, DIF uniforme – não correlacionado e correlacionado positivamente com a dificuldade do item) E tamanho da amostra. | A utilização de um procedimento de correspondência mais "elegante" não conduziu a uma redução dos erros padrão de MH e medidas DIF produzidas eram quase idênticas às obtidas no estudo anterior. Os erros padrão MH tenderam a ser maior quando os itens foram administrados aos respondentes com uma ampla gama de traço latente. |
| Zwick, Thayer e Wingersky (1994a)<br><br>**DIF para itens operacionais do CAT** | MH adaptado para CAT. | Investigar o desempenho de uma versão modificada do método de MH para CAT e comparar com a versão padrão do método *Standardization* para detectar DIF uniforme em itens operacionais. | Tamanho do impacto, tamanho do DIF e tamanho da amostra. | O uso do procedimento adaptado para CAT levou a uma classificação razoavelmente precisa de itens com DIF. Porém, mostrou uma ligeira inflação quando DIF e dificuldade foram positivamente correlacionados. Os resultados não indicaram diferenças consistentes entre os métodos. |

Continuação

| Autor | Método DIF | Objetivo | Variáveis manipuladas | Resultados |
|---|---|---|---|---|
| Zwick e Thayer (2002)<br><br>**DIF para itens operacionais do CAT** | EB DIF; MH | Avaliar, por meio de simulação, a aplicabilidade do método EB DIF em CAT na forma de *testlet.*<br>EB DIF é uma abordagem Bayes empírico do método MH. | Tamanho da amostra e tamanho do impacto. | O método mostrou-se promissor, inclusive para amostras pequenas. No geral, as estimativas EB DIF apresentaram menor RMSE do que as estatísticas MH comuns e obtiveram maiores correlações com os valores-alvo. No entanto, as estimativas EB DIF não são imparciais. Quando EB DIF foi combinado com regra de decisão baseada em função de perda, o método mostrou-se melhor na detecção de DIF do que as abordagens convencionais, mas tem maior taxa de erro Tipo I. |
| Makransky e Glas (2013)<br><br>**DIF entre diferentes contextos PARA CAT** | *Lagrange Multiplier statistic* (LM) e *Wald test* | Apresentar um método para investigar medidas de invariância e modelar o DIF com parâmetros específicos de cada grupo. A estatística LM (baseada em uma tabela de contingência de respostas aos itens por níveis de traço latente em cada subgrupo) é usada para detectar DIF e o teste Wald para significância. | Investiga efeitos de contexto com base no método de aplicação do teste (não-adaptativo *x* CAT) e efeitos devido à linguagem dos respondentes em CATs. | Os resultados forneceram evidências de que a modelagem DIF com parâmetros dos itens específicos de grupo é uma metodologia viável para fazer comparações entre contextos, sem eliminar itens, a não ser por questões legais que impeçam o seu uso. Esses itens devem medir o mesmo construto, embora façam isso de uma maneira diferente em cada grupo. |

Fonte: Elaborado pela autora.

Zwick (2010) apresenta uma revisão sobre métodos para DIF em CAT. Além disso, salienta que diferentes abordagens de simulação são utilizadas para comparar os métodos e avaliam diferentes aspectos, o que torna útil criar um conjunto de dados de simulação comum, sobre o qual poderia ser aplicado a todos os métodos existentes, e um conjunto comum de critérios poderia ser utilizado para avaliar os resultados, tais como o desempenho dos métodos DIF em termos de estimativa dos parâmetros, taxa de erro Tipo I, poder e classificação DIF.

Conforme os estudos apresentados no Quadro 11, fatores como o tipo de DIF, parâmetros dos itens que compõem o CAT, distribuição do traço latente dos respondentes e tamanho da amostra dos grupos de referência e focal podem impactar nos resultados, os quais são comparados, geralmente, pelo poder de detecção do DIF e taxa de erro Tipo I associados.

Não encontrou-se estudos que abordem a análise de DIF quando o método de calibração on-line adaptativo para itens de pré-teste é utilizado. Outros estudos podem ser obtidos em Miller (1992), Walker, Beretvas e Ackerman (2001), Piromsombat (2014).

### 3.1.2 FASE 2 - Monitoramento dos itens

#### 3.1.2.1 Etapa 1 - Monitoramento da exposição dos itens

Em testes de alto impacto, há o risco eminente gerado pelo pré-conhecimento de itens por parte dos respondentes, o qual é agravado devido à superexposição dos itens e sobreposição de testes. Por isso, índices dessas taxas são comumente usados para avaliar a segurança dos testes (CHEN; LEI, 2005; CHEN; LEI; LIAO, 2008, DAVEY, 2011; BARRADA et al., 2009). Esses índices devem ser controlados, uma vez que podem invalidar os resultados de um teste.

Em princípio, quanto mais baixa a taxa de exposição, menor é a quantidade de sobreposição dos itens nos testes (STOCKING, 1994), diminuindo a probabilidade de pré-conhecimento dos itens por alguns respondentes. No entanto, deve haver um equilíbrio entre a precisão e segurança do BI, pois, em geral, os métodos de seleção de itens com maior precisão também são os métodos com os piores índices de exposição (LEROUX et al., 2013).

Alguns estudos analisaram as relações entre as taxas de exposição e sobreposição (BARRADA et al., 2009; CHEN; ANKENMANN;

SPRAY, 2003; CHEN; LEI, 2010), mas conforme destacam Chen et al. (2014), métodos comumente utilizados para controlar a taxa de exposição dos itens, não apresentam mecanismos para controlar conjuntamente a taxa de sobreposição de testes. Nesse sentido, pesquisadores têm se dedicado a desenvolver métodos que visam controlar ambas as taxas (CHEN; LEI, 2005; CHEN; LEI; LIAO, 2008; CHEN, 2010; CHEN et al., 2014), conforme apresentado na seção 2.2.3.3.2.

Para maior segurança, o ideal é utilizar um método de controle de ambas as taxas. Porém, na prática, usualmente utiliza-se um método de controle da taxa de exposição, e a taxa de sobreposição de itens é apenas computada, a qual serve como auxílio para a tomada de decisão sobre qual método de seleção de itens e de controle da taxa de exposição deve ser implantado no algoritmo computacional. Neste processo, estudos simulados avaliam os impactos da inserção do controle dessas taxas na precisão das estimativas dos traços latentes e ajudam nesta decisão.

Uma vez decidido e implantado o método de controle, ele servirá como base do monitoramento dos itens. Esses itens do BI também precisam passar por outras etapas de avaliação, como identificar itens do BI que podem ter se tornado pré-conhecidos e/ou que apresentam DPI. Esses procedimentos servirão como apoio na tomada de decisões durante a manutenção do BI, os quais compõem a FASE 2 da sistemática.

A Figura 9 apresenta os procedimentos gerais abordados na FASE 2 (Etapa 1) da sistemática de manutenção do BI. Assim, após a finalização da edição de aplicação dos CATs, deve-se verificar a taxa de exposição dos itens do BI (ou de ambas as taxas – exposição e sobreposição, quando for o caso). Esta taxa é atualizada automaticamente durante o teste e armazenada.

Itens que atingiram a taxa máxima predeterminada ao final da edição são sinalizados como "**item exposto**" para análise, visando decidir sobre sua exclusão ou permanência no BI. Itens que não atingiram a taxa máxima de exposição, são classificados como **"item não exposto"** e, se necessário, também devem passar por procedimentos de diagnóstico para identificar se há indícios de pré-conhecimento de itens (Etapa 1) e/ou alterações de suas características iniciais/originais (Etapa 2). Essa verificação não é necessária para itens que foram recém calibrados na edição do CAT.

Haverá necessidade de investigação para o pré-conhecimento de **"item não exposto"**, quando o método de controle envolve uma taxa global, visto que um item pode ter sido muito apresentado para

determinado nível do traço latente (tornando-o rapidamente conhecido), mas pouco apresentado em relação à todos os respondentes. Não haverá necessidade de investigação para o pré-conhecimento de **"item não exposto"**, quando o método de controle for condicional ao traço latente. Porém, em ambos os casos, a Etapa 2 deve ser executada.

Desta forma, para um item sinalizado como "**item não exposto"**, verifica-se a necessidade de investigar o pré-conhecimento. Se a resposta for negativa, passa-se para a Etapa 2. Se a resposta for positiva, sugere-se aplicar um método de detecção de itens pré-conhecidos (apresentados a seguir) para auxiliar na decisão de:

- **Opção 1**: não há indícios de pré-conhecimento (item não detectado); logo, o item é mantido no BI e passa-se para a Etapa 2.
- **Opção 2**: há indícios de pré-conhecimento (item detectado); portanto, deve-se investigar os motivos que levaram a isso e ir para a Etapa 2.

Para o **"item não exposto"**, a Etapa 2 objetiva verificar se o item apresenta alterações em seus parâmetros originais e precisa ser eliminado do BI. Isso é necessário, uma vez que: **(1)** o item é "não exposto" e não foi detectado como pré-conhecido, mas pode passar muito tempo no BI sem ser muito utilizado devido à suas características estatísticas, podendo apresentar um conteúdo defasado ao longo do tempo e, consequentemente, *drift* dos parâmetros dos itens; **(2)** o item é "não exposto", mas há indícios de pré-conhecimento. Isso pode ocorrer principalmente nos extremos da escala de medida, onde os respondentes podem receber praticamente os mesmos itens, o que também pode gerar certo *drift* nos parâmetros.

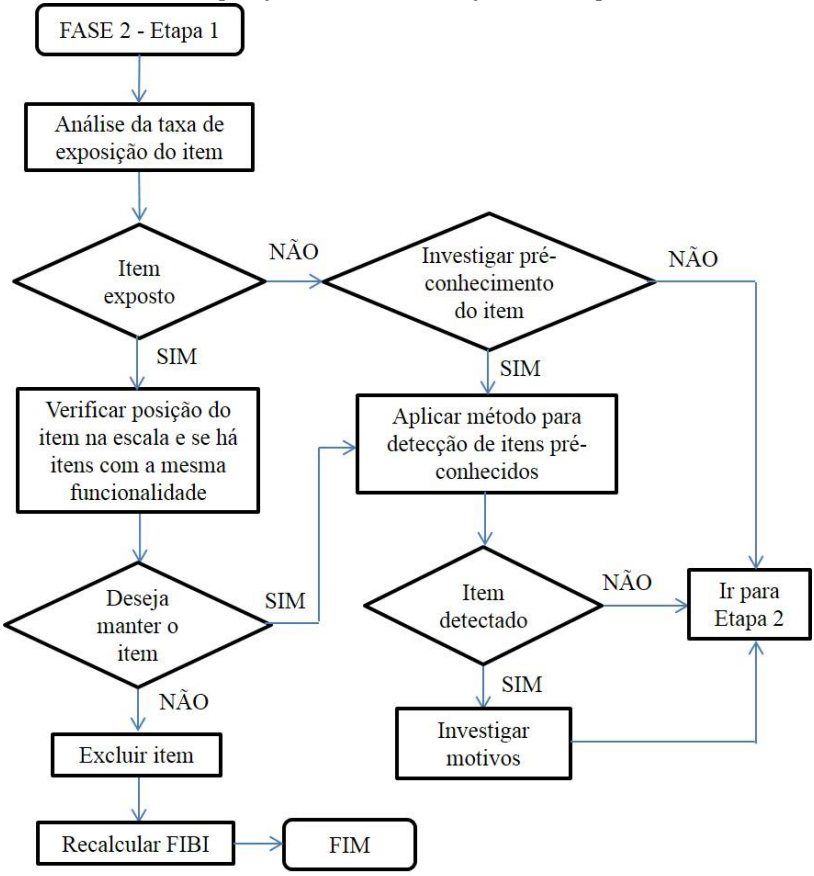
Com base na Figura 9, quando o item é sinalizado como **"item exposto"**, deve-se verificar se há um item no BI com a mesma funcionalidade (em relação ao domínio de conteúdo e aos parâmetros dos itens) e que não tenha atingido a taxa máxima de exposição. Posteriormente, deve-se decidir se é desejável manter o item no BI ou excluí-lo.

A decisão de excluir o item nesta Etapa 1 do processo não é muito indicada, uma vez que elevados custos são gerados com o desenvolvimento de novos itens, e este poderia não estar comprometido, mesmo que tenha atingido a taxa de exposição. Por outro lado, os responsáveis pela manutenção do CAT poderiam optar pela sua exclusão, uma vez que o item tenha atingido esta taxa predefinida e há itens

semelhantes no BI, cujo objetivo desta ação é preservar a integridade do BI, mesmo que isso possa prejudicar em algum grau a avaliação.

Se o item for excluído, deve-se recalcular a FIBI e a Fase 2 é finalizada. Caso deseja-se manter o item que atingiu a taxa máxima de exposição, sugere-se aplicar um método para detectar pré-conhecimento do item e, se detectado, investiga-se os motivos e passa-se para a Etapa 2, a qual também precisa ser executada para itens não detectados.

Figura 9 – Representação da FASE 2 (Etapa 1) de manutenção do BI: monitoramento da exposição dos itens e detecção de itens pré-conhecidos.




Fonte: Elaborada pela autora.

Uma observação importante é que muitos itens podem atingir a taxa máxima de exposição na mesma edição de testes, se o BI não for muito grande. Geralmente, o item frequentemente usado pode ser o "melhor" no BI para a população específica de respondentes (MILLS; STOCKING, 1995), bem como para aquele conteúdo específico.

A eliminação abrupta de muitos itens do BI pode reduzir o alcance efetivo do teste adaptativo (STOCKING, 1988a, 1994; WAY, 2006), arriscando a progressiva deterioração da qualidade global do BI e/ou tornar necessário a prolongação dos testes adaptativos, cujo testes são construídos a partir de um conjunto de menor qualidade (MILLS; STOCKING, 1995).

*3.1.2.1.1 Pré-conhecimento de itens*

Com o aumento de aplicações dos testes adaptativos, preocupações com a segurança dos testes também evoluíram para garantir que um BI seja protegido contra fraudes (MCLEOD; LEWIS; THISSEN, 2003). Essas fraudes podem ocorrer de diversas formas, seja por invasão do sistema ou por respondentes que utilizam alguma estratégia para memorizar os itens que são aplicados e, posteriormente, divulgá-los (ver WISE; KINGSBURY, 2000).

Lu e Hambleton (2004) ressaltam que, independentemente dos controles que são postos em prática pela agência de testes, os itens serão divulgados e é essencial que eles sejam identificados o mais cedo possível para a manutenção da validade do teste. Desta forma, utilizar métodos para tentar identificar itens comprometidos por pré-conhecimento têm ganhado destaque em avaliações de alto impacto.

Yi, Zhang e Chang (2005) desenvolveram o *software AddChart Application* para examinar a relação entre o tamanho do BI, o número de itens que cada "respondente profissional" pode memorizar e a porcentagem de itens que podem ser comprometidos. Logo, este *software* pode auxiliar os profissionais e pesquisadores no desenvolvimento de um CAT mais seguro com base nessas informações.

Ao longo do tempo, vários métodos foram desenvolvidos para detectar comportamentos discrepantes dos respondentes, ou seja, para detectar **respondentes** que podem ter se beneficiado do pré-conhecimento de itens para aumentar seu traço latente em CAT (ver MCLEOD; LEWIS, 1999; MCLEOD; LEWIS; THISSEN, 2003; VAN KRIMPEN-STOOP; MEIJER, 1999, 2001; MEIJER; VAN KRIMPEN-

STOOP, 2010; MEIJER, 2002; – termos usados: *Detecting Person Misfit* ou *Aberrant Response Patterns*).

A abordagem tradicionalmente utilizada é um teste estatístico para resíduos de vetores de resposta (VAN DER LINDEN; GUO, 2008; VAN DER LINDEN; VAN KRIMPEN-STOOP, 2003; QIAN et al., 2016). No entanto, esta técnica tem dificuldade em manter seu poder quando aplicada a CATs porque eles são curtos e, principalmente, porque a dificuldade dos itens convergem para o nível do traço latente do respondente (VAN DER LINDEN; VAN KRIMPEN-STOOP, 2003).

Sendo assim, itens próximos ao final do teste tendem a ter uma probabilidade de sucesso perto de 0,50, sendo que a análise residual normalmente tem sua potência máxima para testes com probabilidades de respostas corretas próximas de um ou zero; logo, tem-se uma baixa taxa de detecção e uma alta taxa de falsos alarmes (VAN DER LINDEN; GUO, 2008; VAN DER LINDEN; VAN KRIMPEN-STOOP, 2003; QIAN et al., 2016).

Devido às sérias críticas de acusar alguém de ser um usuário do pré-conhecimento do item (ou "trapaceiro"), pode ser mais útil para os desenvolvedores de testes se concentrarem na segurança do item, em vez do comportamento dos respondentes individuais (MCLEOD; SCHNIPKE, 1999). Conforme Lu e Hambleton (2004), detectar itens que estão se comportando de forma anormal por causa da divulgação é, em alguns aspectos, semelhante a identificação de respondentes que dão padrões de resposta discrepantes. Nessa perspectiva, alguns métodos estatísticos e índices para detectar **itens** comprometidos foram desenvolvidos.

*3.1.2.1.2 Métodos para detectar pré-conhecimento*

Alguns trabalhos serão apresentados como sugestões de métodos para identificação do pré-conhecimento de itens, os quais apresentam duas abordargens distintas, uma com base no tempo de resposta e outra baseada em estatísticas.

**I. Tempo de resposta ao item para detectar itens comprometidos**

van der Linden e Guo (2008) apresentam três importantes razões pelas quais os tempos de resposta (RTs) podem ser uma melhor fonte de informação sobre possíveis inconsistências do que as próprias respostas:

**1º)** Os dados são contínuos em vez de binários, proporcionando melhor tratamento estatístico e análise do "tamanho" das inconsistências quando elas acontecem;

**2º)** Os RTs são insensíveis ao efeito adaptativo dos testes, ou seja, se um teste coincide com o nível do traço latente do respondente, a seleção dos itens não restringe a probabilidade de suas RTs de forma sistemática e continua sendo possível discriminar entre padrões prováveis e improváveis de RTs, mantendo o poder de testes estatísticos baseados em RT durante todo o teste;

**3º)** Os RTs são o resultado da velocidade com que os respondentes trabalham nos itens, bem como suas intensidades de tempo, a qual é determinada pela quantidade de trabalho que ele requer para produzir uma resposta. Não importa o quão rápido ou lento um respondente trabalha, os RTs têm que seguir o padrão de intensidades de tempo dos itens no teste, se este for bem projetado.

Logo, se há um modelo de RT que permite separar a velocidade das intensidades de tempo, é possível ajustar os RTs dos respondentes para a sua velocidade e verificar se os resultados seguem o padrão de intensidades de tempo. Mesmo que os respondentes simulassem RTs para esconder uma fraude, eles teriam que descobrir qual é o padrão típico para os itens que eles recebem, o que é praticamente impossível, especialmente porque ele deve ser executado conforme o tempo que já está gravado (dados da calibração, por exemplo) (VAN DER LINDEN; GUO, 2008).

Qian et al. (2016) reiteram que os respondentes não falsificarão o tempo de resposta, sendo esta uma suposição realista por duas razões:

**1º)** A maioria dos respondentes não sabe que o seu RT é monitorado e usado para análise porque a maioria dos relatórios de pontuação não mostram essa informação;

**2º)** Para os respondentes que têm pré-conhecimento do item, eles não só querem responder o item atual corretamente, mas também rapidamente, para que eles possam ter mais tempo para outros itens, uma vez que a maioria dos testes são de tempo limitado. Assim, após identificar um item familiar e recordar uma resposta memorizada, eles iriam passar para o próximo item e economizar mais tempo para outras perguntas onde eles não têm pré-conhecimento e que refletirão o tempo realista.

Desse modo, o RT pode facilmente ser armazenado e, particularmente em testes de alto impacto, podem refletir o tempo realmente necessário para o respondente processar os itens e produzir uma

resposta realista, podendo ser usado como suporte para verificação de inconsistências (VAN DER LINDEN; VAN KRIMPEN-STOOP, 2003), as quais podem apresentar tipos específicos de comportamento, conforme segue:

- **Quando o RT é inferior ao esperado (respostas rápidas) e a resposta é <u>incorreta</u>** - pode ser devido à resposta dada aleatoriamente (WISE, 2014; VAN DER LINDEN; VAN KRIMPEN-STOOP, 2003).

Neste caso, o ideal é verificar a posição desses itens no teste, buscando identificar possíveis motivos do ocorrido. Este comportamento pode indicar descomprometimento com o teste ou, então, se o mesmo estiver localizado no final do teste, pode indicar falta de tempo para completar o teste, fazendo que o respondente dê respostas aleatórias para encerrar o teste dentro do prazo; a fadiga também pode ser um dos motivos em testes extensos.

Wise (2014) afirma que as respostas rápidas são aquelas que ocorrem mais rapidamente do que deveriam ser esperadas para um respondente ao ler, compreender e selecionar a resposta; também, como o CAT aplica itens direcionados ao nível do traço latente dos respondentes, um conjunto de respostas com uma baixa taxa de precisão pode fornecer evidências adicionais para identificar o baixo esforço do respondente.

- **Quando o RT é inferior ao esperado (respostas rápidas) e a resposta é <u>correta</u>** - pode ser devido à resposta dada aleatoriamente ou devido ao pré-conhecimento do item (VAN DER LINDEN; VAN KRIMPEN-STOOP, 2003; VAN DER LINDEN; GUO, 2008; QIAN et al., 2016). Este resultado pode acontecer para itens em qualquer posição no teste (VAN DER LINDEN; VAN KRIMPEN-STOOP, 2003).

Por outro lado, existem fatores que fazem com que o **RT seja muito superior ao esperado**:

- Pode evidenciar que o item é ruim quando este comportamento ocorre para muitos respondentes (VAN DER LINDEN; VAN KRIMPEN-STOOP, 2003; VAN DER LINDEN; GUO, 2008);
- Estudos mostram que itens mais difíceis geralmente requerem mais tempo dos respondentes (para respondentes com elevado traço latente) (VAN DER LINDEN; SCRAMS; SCHNIPKE, 1999; CHANG; PLAKE; FERDOUS, 2005);

- Tentativa de memorização de itens para posterior divulgação a outros candidatos. Pode haver uma sequência de longo tempo e respostas incorretas, isso porque em CAT é necessário inserir uma resposta para fazer o próximo item aparecer na tela (VAN DER LINDEN; GUO, 2008).

Nota-se que existem vários padrões de inconsistências no RT, que podem ocorrer por vários motivos. Desta forma, sua identificação nem sempre é fácil de ser justificada. Por isso, van der Linden e Van Krimpen-Stoop (2003) enfatizam que é prudente considerar o controle das respostas e dos RTs apenas como apoio a uma hipótese de fraude e não como prova decisiva. Com base nisso, sugere-se obter um conjunto de evidências, passando pelas etapas 1 e 2 da FASE 2, para então decidir sobre o destino final do item.

**Uso de um modelo para RTs e representação gráfica na detecção**

O método aqui apresentado foi utilizado nos estudos de Qian et al. (2016) e van der Linden e Guo (2008), apresentando uma boa taxa de detecção. Eles utilizaram uma estrutura hierárquica para detectar padrões discrepantes de RTs. Para representar os resultados, o método faz uso de um procedimento gráfico baseado em resíduos padronizados do logarítmo do tempo de resposta para identificar itens comprometidos. O método também pode ser usado para detectar comportamentos inconsistentes dos respondentes.

Segundo van der Linden e Guo (2008), a estrutura hierárquica visa analisar a velocidade e a precisão nos itens, ou seja, usa uma combinação do modelo para RT com um modelo de resposta da TRI em uma estrutura hierárquica que permite usar as informações do vetor de respostas como informações colaterais sobre a velocidade do respondente em uma verificação bayesiana.

- **Validação dos modelos** (QIAN et al., 2016)

Para utilizar os modelos para prever RTs razoáveis, compará-los com os observados e identificar os inesperados, é necessário primeiro testar se os dados se ajustam aos modelos. **1º)** Se o modelo da TRI se ajusta aos itens; **2º)** Se os RTs correspondem à distribuição lognormal, cujo modelo foi proposto por van der Linden (2006).

Conforme van der Linden e van Krimpen-Stoop (2003), o modelo lognormal foi comparado aos modelos baseados nas distribuições Normal, *Gamma* e *Weibull* e mostrou excelente ajuste para RTs. Portanto,

se ambos os modelos se ajustam, respostas e RTs para um determinado respondente são independentes e as estatísticas de teste definidas nessas variáveis também são independentes (VAN DER LINDEN; VAN KRIMPEN-STOOP, 2003).

- **Procedimentos para detectar o pré-conhecimento** (QIAN et al., 2016)

Para detectar o pré-conhecimento do item são necessárias duas amostras. A primeira amostra inclui uma matriz das respostas e uma matriz dos RTs registrados na calibração do item, ou seja, fase inicial onde não há itens comprometidos. A segunda amostra é obtida a partir da fase posterior do teste operacional, que está sujeita à problemas de exposição de conteúdo, ou seja, itens que possam estar comprometidos.

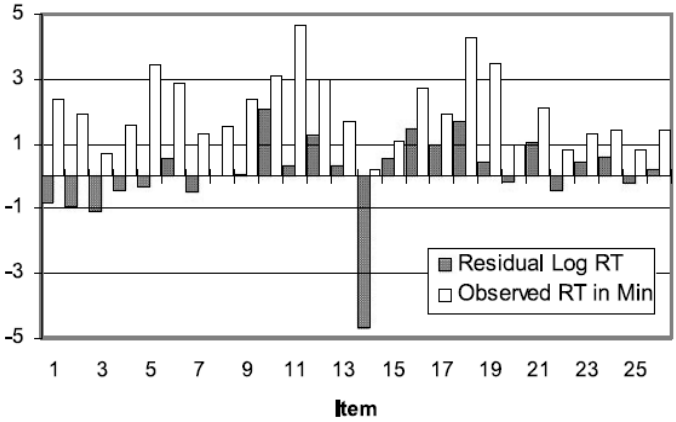
Qian et al. (2016) e van der Linden e Guo (2008) analisaram apenas itens operacionais com pelo menos 100 respostas. O RT inesperadamente curto é detectado pelo resíduo do log do RT estimado. Tais resíduos têm uma distribuição empírica aproximada N(0,1) quando itens não estão comprometidos.

Como os parâmetros de velocidade são estimados para o teste completo, um aumento na velocidade real em subconjuntos de itens se manifesta por valores negativos maiores para os resíduos. Assim, um RT curto a um item é sinalizado como inconsistente quando seu residual tem um valor negativo menor do que -1,96, conforme pode ser observado na Figura 10. Valores acima de +1,96 são sinalizados como inconsistente porque o respondente gastou mais tempo no item do que o esperado com base na velocidade dele e na intensidade de tempo do item.

Para **detectar indivíduos suspeitos**, obtém-se um resumo da frequência com que os RTs para o mesmo respondente é sinalizado. Para **identificar itens suspeitos**, verifica-se a frequência com que os RTs do mesmo item são marcados como suspeitos e se há qualquer padrão entre o tamanho dos resíduos (VAN DER LINDEN; GUO, 2008).

Deve-se monitorar os RTs ao longo do tempo. Assim, os resíduos para todos os respondentes nas duas amostras são comparados para verificar se houve mudanças (QIAN et al., 2016). Geralmente, levará algum tempo para um item ficar comprometido após o seu lançamento; antes que isso aconteça, os resíduos serão regulares; após algum tempo, pode-se esperar que mostrem uma tendência para uma média mais baixa (VAN DER LINDEN; GUO, 2008).

Figura 10 – Representação da análise dos RTs para identificação de itens pré-conhecidos, com o item 14 apresentando resíduo menor do que -1,96.




Fonte: van der Linden e Guo (2008). Nota: na cor cinza tem-se os resíduos do log RT e na cor branca, para os RTs observados (em minutos).

Nesta linha, van der Linden e Guo (2008) recomendaram a análise de três indicadores de pré-conhecimento de um item por um respondente. Isto é, além de uma resposta correta e um RT residual negativo grande, observar uma baixa probabilidade de sucesso no item (ou seja, um baixo traço latente estimado ao final do teste em relação à dificuldade do item).

No estudo de van der Linden e Guo (2008), houve um total de quatro respondentes com o mesmo tipo de registro para o item 14 (Figura 10). Ou seja, foi gasto um tempo muito curto no item (aproximadamente 12-13 segundos, enquanto o item exigia aproximadamente 88,9 segundos), obteve-se uma resposta correta e traço latente estimado muito menor que o parâmetro *b* do item. Porém, o autor conclui que é necessário um número muito maior de respondentes com este comportamento no item para concluir que existe uma ameaça real à segurança do teste.

van der Linden e Guo (2008) utilizaram três períodos de testes. Para cada período foram calculadas as médias e desvios padrão das RTs residuais em todos os itens. As médias dos RTs residuais de cada período foram muito próximas. A ANOVA (análise de variância) unidirecional não mostrou diferença significativa entre os períodos analisados. De acordo com os autores, uma alternativa à abordagem ANOVA seria utilizar estatísticas baseadas em CUSUM (procedimentos de soma

cumulativa) sobre RTs para itens individuais coletados ao longo do tempo.

Como as estatísticas CUSUM tiveram origem da análise de séries temporais para fins de controle de qualidade, elas podem ser utilizadas para verificar se séries de RTs em itens individuais em CAT permanecem dentro de limites de qualidade (VAN DER LINDEN; GUO, 2008).

**II. Estatísticas para detectar pré-conhecimento de itens**

Nesta linha de pesquisa, apresenta-se o trabalho de Lu e Hambleton (2004) como opção de estatísticas direcionadas à detecção de itens divulgados e que se tornam pré-conhecidos. Os autores propõem duas novas estatísticas (uma baseada em uma análise residual e a outra baseada em uma análise da função de verossimilhança com e sem a divulgação de itens), investigam as distribuições nulas das estatísticas propostas em um ambiente CAT e definem os limites quando essas distribuições não possuem propriedades conhecidas. Por fim, eles determinam a eficácia das estatísticas na detecção de itens divulgados por meio de um estudo de simulação.

Segundo os autores, a primeira estatística ($K_1$) é um resíduo padronizado, citado em Zhu, Yu e Liu (2002) para o monitoramento sequencial do desempenho do item em um CAT. Esta estatística está relacionada ao número total de respondentes obtendo respostas corretas para um item, comparando-a com o número esperado de respostas corretas com base na distribuição do traço latente do grupo de respondentes, assumindo ajuste de dados ao modelo. Para itens divulgados, $K_1$ vai assumir valores falsamente elevados.

$$K_1 = \frac{\sum_{j=1}^{J}\left(x_{ij} - P_{ij}\right)}{\sqrt{\sum_{j=1}^{J} P_{ij}\left(1 - P_{ij}\right)}} \qquad (13)$$

onde $x_{ij}$ é o traço latente binário da variável para o item $i$ do respondente $j$; $P_{ij}$ é a probabilidade que o respondente $j$, com traço latente $\theta_j$, dê uma resposta correta para o item $i$; $J$ é o número de indivíduos que responderam ao item $i$ recentemente.

Na prática, o verdadeiro traço latente do respondente não é conhecido; logo, $P_{ij}$ não é alcançável. Portanto, os autores propõem a estatística $*K_1$, a qual usa o traço latente estimado do respondente, bem como a estimativa $\hat{P}_{ij}$.

A segunda estatística ($K_2$) é uma função log-verossimilhança padronizada e é uma variação da estatística *person-fit lz* proposto por Drasgow, Levine e Williams (1985). De acordo com Lu e Hambleton (2004), $l_z$ é calculado ao nível do respondente para detecção de padrões de respostas discrepantes, enquanto $K_2$ é calculado no nível do item para a detecção de itens discrepantes. $K_2$ difere de $l_z$ na medida em que examina a log-verossimilhança das respostas observadas de um grupo respondentes para um item, em vez da log-verossimilhança de respostas observadas de um respondente a um conjunto de itens.

Assim, esta estatística compara a verossimilhança dos dados observados com a verossimilhança esperada dos padrões de resposta sob o modelo assumido. A divulgação do item conduzirá a uma pequena log-verossimilhança normal padronizada dos padrões de resposta observados. $K_2$ é dado por:

$$K_2 = \frac{l - E(l)}{\sqrt{Var(l)}} \tag{14}$$

onde $l$ é a log-verossimilhança do padrão de respostas observadas de $J$ respondentes ao item $i$, E($l$) e Var($l$) são a esperança e a variância de $l$, respectivamente, dados por:

$$l = \sum_{j=1}^{J} \left\{ X_{ij} \ln P_{ij} + \left(1 - X_{ij}\right) \ln\left[1 - P_{ij}\right] \right\} \tag{15}$$

$$E(l) = \sum_{j=1}^{J} \left\{ P_{ij} \ln P_{ij} + \left(1 - P_{ij}\right) \ln\left[1 - P_{ij}\right] \right\} \tag{16}$$

$$Var(l) = \sum_{j=1}^{J} Var\left\{ X_{ij} \ln P_{ij} + \left(1 - X_{ij}\right) \ln\left[1 - P_{ij}\right] \right\} = \sum_{j=1}^{J} P_{ij}\left[1 - P_{ij}\right] \left[ \ln \frac{P_{ij}}{1 - P_{ij}} \right]^2 \tag{17}$$

Novamente, como em $K_1$, $P_{ij}$ em $K_2$ deve ser substituído por $\hat{P}_{ij}$, já que o verdadeiro traço latente não é conhecido. Portanto, a nova estatística utilizando o traço latente estimado é denotada como $*K_2$.

Neste estudo de Lu e Hambleton (2004) foi considerada uma amostra de 452 itens que haviam sido respondidos por 400 indivíduos. Na prática, a escolha do tamanho da amostra depende do tamanho da população de respondentes por ano. Quando o tamanho da população anual é pequena, um ponto de equilíbrio deve ser encontrado entre o poder estatístico e sensibilidade.

$K_1$ é capaz de detectar somente desvio uniforme nos dados do modelo e $K_2$ deve ser capaz de identificar qualquer tipo de padrões de resposta desajustados. No entanto, os autores destacam que não foram gerados dados deste tipo no estudo para detecção de desvio não-uniforme. Testes como Shapiro-Wilk e *p*-valor foram efetuados para verificar a normalidade dos dados.

Como resultado, os autores identificaram que as estatísticas $K_1$ e $*K_1$ têm distribuição normal padronizada e o quantil 95% de uma distribuição normal padrão foi usado como critério para a sinalização de itens divulgados (valor crítico de 1,645). Já para as estatísticas $K_2$ e $*K_2$, a hipótese de distribuição normal foi rejeitada, por isso, o quantil 5% da distribuição simulada de $*K_2$ (-0,718) foi usado como o valor de corte.

Verificou-se que $K_1$ e $*K_1$ tiveram taxas de detecção mais elevadas, enquanto as taxas de detecção de $K_2$ e $*K_2$ foram moderadas. Na prática, as estatísticas $K_1$ e $K_2$ não podem ser computadas porque os traços latentes verdadeiros não são conhecidos. Então, sob as condições simuladas, as principais conclusões são que $*K_1$ é assintoticamente distribuída como N(0,1) e obteve uma taxa de detecção de 72%, alcançável com uma taxa de erro tipo I de 5%. Embora $*K_2$ não tenha uma distribuição nula identificável e apresentou uma moderada taxa de detecção de 0,26, que é muito mais baixa do que a taxa de $*K_1$, teoricamente, pelo menos, $*K_2$ é capaz de identificar os desvios não-uniformes e uniforme no modelo de dados e, portanto, também pode ser útil na prática. Assim, os autores concluíram que ambas as estatísticas investigadas parecem promissoras.

#### 3.1.2.2 Etapa 2 – Verificação de *drift* dos parâmetros dos itens

A invariância dos parâmetros dos itens é crucial quando se quer avaliar e comparar populações de respondentes ou condições de medição. Porém, podem ocorrer alteração nos parâmetros dos itens em testes subsequentes, ou seja, *drift* dos parâmetros do item (DPI).

Problemas que surgem ao longo do tempo podem ser diagnosticados por meio da manutenção do BI (DONOGHUE; ISHAM,

1998; THISSEN et al., 2007). Deste modo, os itens devem ser revisados periodicamente por profissionais especializados para garantir que eles são atuais e relevantes para o campo da prática e para assegurar a validade dos resultados de um teste (BERGSTROM; GERSHON, 1995; WAY, 2006; SQUIRES, 2003; WALKER et al., 2010; HAN; GUO, 2011; CLARK, 2013). Esses profissionais podem ajudar a identificar as causas do DPI.

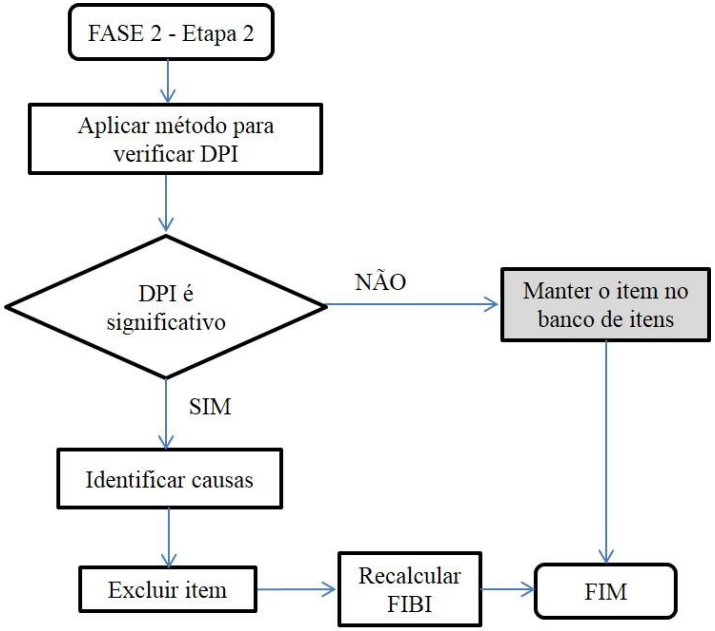
Na Etapa 1, quando um item é detectado como pré-conhecido, é provável que ele exiba DPI. Por isso, esta etapa da sistemática dará suporte para confirmar se isso realmente ocorreu. Quando um item é apresentado muitas vezes, os respondentes podem se familiarizar com ele e podem se preparar para isso (pré-conhecimento do item), o que resultaria numa diminuição da dificuldade real do item (*drift*), que por sua vez, gera um viés positivo na estimativa do traço latente (GEORGIADOU; TRIANTAFILLOU; ECONOMIDES, 2007).

Se um item não atingiu a taxa máxima de exposiçaõ e não apresenta indícios de pré-conhecimento, também precisa passar pela verificação do DPI, pois existem outros motivos que podem levar ao *drift*. Por exemplo, ao longo do tempo, itens tornam-se obsoletos e/ou passam a funcionar mal (após serem avaliados estatisticamente), sendo necessário substituí-los (SQUIRES, 2003).

Independente da causa do DPI, sua existência representa uma ameaça para testes que exigem uma escala estável para mensurar o traço latente e precisam ser identificados (WELLS; SUBKOVIAK; SERLIN, 2002; GUO; WANG, 2003; WELLS et al., 2014). Portanto, para dar sequência ao monitoramento dos itens e auxiliar nas decisões relacionadas ao DPI, elaborou-se a Fase 2 (Etapa 2), conforme mostra a Figura 11.

Conforme a Figura 11, primeiramente aplica-se um método para detectar DPI. Caso o item não apresente DPI significante, ele pode ser reutilizado, mantendo-o no BI e finalizando a Etapa 2. Se o item apresentar DPI significativo, busca-se identificar as causas (de modo a tentar evitar, no futuro, novos casos) e sugere-se excluir o item. Novamente, a FIBI deve ser recalculada, finalizando esta Etapa 2 de manutenção do BI para a edição de aplicação de CATs.

Figura 11 – Representação da FASE 2 (Etapa 2) de manutenção do BI: verificação de *drift* dos parâmetros dos itens.




Fonte: Elaborada pela autora.

### *3.1.2.2.1 Métodos para detecção de drift dos parâmetros do item – DPI*

Na literatura existem inúmeros métodos desenvolvidos para detectar *drift* em testes tradicionais (ver DONOGHUE; ISHAM, 1998; RAJU, 1988; DEMARS, 2004). No entanto, poucos estudos têm sido desenvolvidos no contexto de DPI em CAT (GUO, 2016; LU; HAMBLETON, 2004; RISK, 2015).

Tipicamente, os itens que exibem DPI são identificados por comparação dos parâmetros do item ou da CCI entre os respectivos grupos (WELLS et al., 2014). Segundo Guo (2016), o uso de CCI para a detecção de DPI é limitado devido ao efeito de amplificação, que mostra DPI óbvio a nível total (FIT) quando os itens sofrem mudanças para a mesma direção e o efeito de cancelamento quando dois conjuntos de itens apresentam DPI em direções opostas, deixando a FIT sem *drift*.

Muitos autores sugerem como método para detectar DPI a reestimação ou recalibração dos parâmetros dos itens rotineiramente em CAT, comparando-as com as calibrações iniciais (VEERKAMP; GLAS, 2000; LU; HAMBLETON, 2004; GUO, 2016; STOCKING, 1988b). Lu e Hambleton (2004) destacam que para isso, se assume que as estimativas obtidas nas duas calibrações são precisas; porém, a precisão pode não ser possível para a recalibração com dados do CAT operacional, uma vez que isso requer uma amostra razoavelmente grande e heterogênea, ou então, que essa condição pode ser atingível para itens administrados no início de um CAT, mas não são atingíveis para os itens administrados mais tarde, quando os itens são próximos ao nível do traço latente.

Outra dificuldade encontrada para a recalibração de itens é a matriz de resposta incompleta gerada. Conforme Harmes, Parshall e Kromrey (2003), esta matriz incompleta de dados é o resultado do tamanho do BI em relação ao comprimento médio do CAT e ao método de seleção de itens utilizado, que leva em consideração o traço latente estimado; logo, a razão para a dispersão na matriz de dados é não-aleatória.

Com isso, nem todos os itens com *drift* podem ser recalibrados de forma adequada, uma vez que alguns dos itens podem ser respondidos por uma proporção muito pequena de respondentes e pode levar a imprecisões na estimação dos parâmetros, prejudicando a qualidade de detecção de DPI (GUO, 2016; HARMES; PARSHALL; KROMREY, 2003). Também, pode levar um longo período de tempo para acumular um número suficientemente grande de respondentes com uma ampla gama do traço latente, o que é bastante inconveniente, e às vezes impossível em CAT, necessitando de abordagens mais práticas para a detecção precoce de itens com DPI (ZHU; YU; LIU, 2002).

Portanto, estratégias para lidar com dados não-aleatórios oferecem possíveis soluções para estes problemas de dados incompletos (HARMES; PARSHALL; KROMREY, 2003). Segundo Guo (2016), métodos mais antigos de detecção de DPI usam calibração de itens separados e um método de ligação para detectar itens com *drift*, apresentando várias limitações.

Atualmente, a nova abordagem de calibração adaptativa on-line, que é usada para calibrar itens de pré-teste, também pode ser usada para recalibrar itens existentes e detectar DPI (GUO, 2016; ZHENG, 2014). Uma grande vantagem desta, é que os parâmetros do item calibrado estão automaticamente na mesma escala existente (ZHENG, 2014). A seguir,

serão apresentados alguns estudos e métodos para análise de DPI no contexto de CAT.

**I. Recalibração não-adaptativa on-line dos parâmetros**

Segundo Ito e Sykes (1994), uma maneira óbvia de recalibrar os itens seria a recalibração não-adaptativa on-line, ou seja, os itens previamente marcados e que serão testados, devem ser tratados como itens de pré-teste. Desta forma, eles serão (quasi-) aleatoriamente administrados aos respondentes, obtendo assim, respondentes com uma ampla gama do traço latente. No entanto, esta forma gera custos adicionais às operações normais de um CAT e, para evitar esses custos extras, a recalibração pode usar os dados obtidos a partir do CAT operacional.

Nesta linha, os autores investigaram por meio de simulações a recalibração utilizando níveis do traço latente próximos ao nível de dificuldade do item para o modelo de Rasch, isto é, itens mais difíceis são calibrados com respostas de indivíduos mais capazes e itens mais fáceis calibrados com respostas de indivíduos menos capazes. Este cenário é obtido em um CAT operacional.

Sob as condições investigadas em Ito e Sykes (1994), os resultados mostraram que os valores dos parâmetros de dificuldade não foram estimados adequadamente quando itens difíceis foram calibrados usando respostas de indivíduos capazes e itens fáceis foram calibrados usando respostas de indivíduos menos capazes, necessitando de modificações na amostra para obter uma melhor calibração dos parâmetros.

Neste contexto, os resultados foram razoavelmente bons apenas para itens e respondentes no nível médio. Portanto, os autores sugerem: (1) que uma amostra de recalibração CAT razoavelmente grande seja predefinida a partir do grupo de referência; (2) a amostra tenha um traço latente médio semelhante à da grupo de referência; e (3) itens a serem recalibrados juntos são relativamente heterogêneos na dificuldade.

**Técnica de imputação de dados para recalibração**

Harmes, Parshall e Kromrey (2003) investigaram a eficácia relativa dos tratamentos de dados faltantes aplicados às matrizes obtidas de administrações CAT sob uma variedade de condições de comprimento do teste, tamanho da amostra e algoritmos de seleção de itens. Os dados faltantes referem-se à matriz incompleta de respostas que são obtidas em CAT.

Eles utilizaram dois métodos: estimativa de máxima verossimilhança com algoritmo EM e imputação múltipla (ver THOMAS; GAN, 1997). As condições investigadas no estudo representaram valores médios de 83% a 92% de dados faltantes nas matrizes de resposta ao item. O *software* Bilog foi utilizado nas análises.

Segundo os autores, o Bilog não foi capaz de calibrar com sucesso as matrizes de dados de resposta incompleta, sem tratamentos dos dados, que pode ser devido a itens individuais para os quais o Bilog não teria dados suficientes para a calibração, impactando em todo o conjunto de dados. Uma alternativa seria utilizar somente um subconjunto dos itens com dados suficientes para calibração no Bilog.

No entanto, esta solução não resolve o problema da calibração de conjuntos de dados do CAT. Em geral, a técnica de imputação múltipla mostrou-se uma ferramenta promissora no contexto da calibração de itens em CAT, com níveis relativamente pequenos de viés e EP e desempenhou melhor nessas estatísticas em relação ao método MV com o algoritmo EM. Variações no comprimento do teste levaram a alterações muito pequenas no EP médio, bem como o tamanho da amostra. As variações no comprimento do teste tiveram um pequeno efeito no viés, diferentemente do tamanho da amostra, que teve maior efeito.

**II. Recalibração adaptativa on-line**

Guo (2016) apresentou um *design* de recalibração adaptativa on-line, centrado no item, para detectar o DPI para CAT unidimensional e multidimensional. Um *design* de dois estágios modificado é proposto para CAT unidimensional pela implementação de um algoritmo de índice de densidade proporcional (*Proportional Density Index* - PDI).

O algoritmo PDI é uma nova estrutura para a seleção de itens de pré-teste. A ideia principal é estabelecer uma dificuldade limite calculada a partir da área do modelo da TRI. A distância entre este limite e o local do respondente é utilizada para decidir se o item de pré-teste candidato é então administrado ao respondente atual. Os estudos foram conduzidos pelo autor via programa Matlab.

O *design* de recalibração adaptativa para detectar DPI é mostrado na Figura 12 e deve seguir as etapas abaixo:

**Etapa 1:** *Separações do BI e do teste*

- Dividir o BI em BI operacional e de recalibração: o BI operacional contém itens que se acredita não ter sofrido *drift*, enquanto o BI de recalibração contém itens suspeitos. Por

exemplo, itens muito expostos ao longo de um tempo, itens que foram diagnosticados como pré-conhecido na Fase 2 (etapa 1) da sistemática ou itens que não foram muito aplicados, mas já estão no BI há bastante tempo e podem ter um conteúdo defasado.

- Dividir o CAT em teste operacional e teste de recalibração: por exemplo, o CAT terá comprimento fixo de 27 itens e mais 3 itens para recalibração. O CAT operacional tem por objetivo estimar o traço latente dos respondentes e seus itens são selecionados a partir do BI operacional usando um método de seleção de itens, como MFI, etc. No teste de recalibração, os itens são selecionados do conjunto de recalibração de acordo com outro método, como o algoritmo PDI, cujo objetivo desse processo é selecionar respondentes adequados para recalibrar os parâmetros do item.

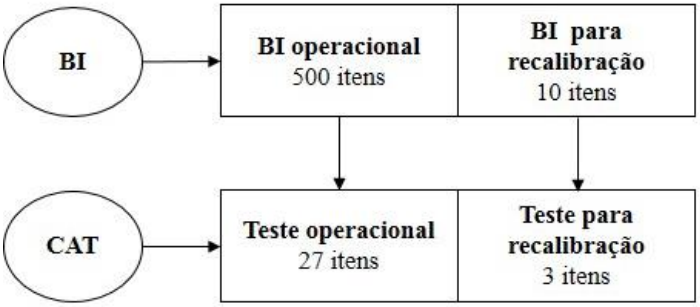
**Etapa 2:** *Estimativa do traço latente e recalibração de itens*

- Estimar o traço latente dos respondentes com base nas respostas aos itens operacionais. Ex.: método EAP;
- Realizar a recalibração on-line: nesta etapa é preciso inicializar os parâmetros do item de pré-teste, definir o método de seleção de itens (ex. PDI) e o local de inserção dos itens de pré-teste, definir o método de estimação e a regra de parada para a recalibração do item, ou seja, quando deve-se terminar a amostragem. Cada vez que um item de pré-teste atingir um critério de parada predeterminado, o processo é concluído.
- Realizar testes de hipóteses como a razão de verossimilhança (THISSEN; STEINBERG; WAINER, 1993), NCDIF (*non-compensatory diferential item functioning method* – RAJU; VAN DER LINDEN; FLEER, 1995), entre outros, para detectar se houve *drift* e então, excluir o item. Outra opção é atualizar os valores dos parâmetros dos itens detectados e mantê-los no BI como um novo item.

No cenário avaliado por Guo (2016), as seguintes conclusões foram obtidas: (1) a técnica de calibração adaptativa é a maneira mais desejável para detectar DPI em comparação com a calibração tradicional com a matriz incompleta, obtida a partir do CAT, pois mais itens podem ser recalibrados e, portanto, maior eficiência de calibração e precisão de detecção DPI pode ser obtida; (2) sob o modelo da TRI unidimensional, o *design* de dois estágios com o algoritmo PDI é o método mais eficiente

em termos de viés, RMSE, taxa de erro tipo I e poder. Este método foi comparado com a seleção aleatória, comparação direta dos valores de D-*optimal* e utilizando o *Suitability index* – SI (ALI; CHANG, 2014); (3) dentre os três métodos de estimação utilizados, a estimativa de máxima verossimilhança condicional tem a menor eficiência de estimativa e as duas versões bayesianas de OEM e MEM tiveram desempenho semelhante.

Figura 12 – *Design* de recalibração on-line para detectar DPI.




Fonte: Adaptado de Guo (2016).

## III. Outras estatísticas de detecção do DPI

De acordo com Veerkamp e Glas (2000), os testes de DPI são casos especiais de testes de viés de item ou de DIF, onde um grupo de referência é comparado com o comportamento de grupos focais; já em estudos de DPI, pode-se distinguir uma fase de calibração de uma outra fase para avaliar se o comportamento de resposta difere ou não.

Nesse contexto, alguns estudos para CAT foram desenvolvidos. Veerkamp e Glas (2000) sugerem o uso do método CUSUM para detectar itens conhecidos devido à divulgação e o seu efeito nos parâmetros dos itens. Este caso implica em um teste unilateral, em que o item se torna mais fácil e perde seu poder de discriminação, pois devido ao conhecimento prévio do item, a probabilidade de resposta correta aumenta. O método não serve para detectar *drift* em direções opostas.

Segundo os autores, o método é baseado no teste sugerido em Lord (1980), em que a a hipótese nula $\hat{b}_i^1 - \hat{b}_i^0 \geq 0$ é testada contra a alternativa $\hat{b}_i^1 - \hat{b}_i^0 < 0$, onde $\hat{b}_i^0$ é o valor do parâmetro de dificuldade na fase de

calibração inicial e $\hat{b}_i^1$ é o valor do parâmetro a ser comparado para DPI na fase posterior do CAT. Assumindo que $b_i^0$ é calibrado (talvez recalibrado após a utilização do item no CAT) como $\hat{b}_i^0$, com erro padrão $\sigma\left(\hat{b}_i^0\right)$, a estatística de teste é dada por:

$$\frac{\hat{b}_i^0 - \hat{b}_i^1}{\sqrt{\sigma^2\left(\hat{b}_i^0\right) + \sigma^2\left(\hat{b}_i^1\right)}} \tag{18}$$

onde $\hat{b}_i^1$ e $\sigma\left(\hat{b}_i^1\right)$ são as estimativas do parâmetro de dificuldade e seu EP com base nos dados coletados durante a administração CAT. Como $\hat{b}_i^0$ e $\hat{b}_i^1$ são estimadas usando amostras independentes, as estimativas não covariam e o EP da diferença $\hat{b}_i^0 - \hat{b}_i^1$ pode ser calculado como o denominador da Equação 31. A estatística é conhecida como teste do tipo Wald e tem uma distribuição assintótica normal padronizada.

Segundo os autores, o procedimento CUSUM pode ser visto como uma série sequencial de testes Wald, onde o teste continua sempre até que a hipótese nula de nenhuma mudança é rejeitada. O poder do teste é uma função do número de amostras. Para o controle de qualidade de itens em um BI CAT, este método pode ser baseado em desvios acumulados de estimativas de parâmetros de dificuldade a partir do valor encontrado no estudo de calibração.

Os parâmetros dos itens foram reestimados usando MML. Os resultados mostraram que a taxa de detecção deste procedimento é bastante aceitável e que a taxa de erro tipo I está sob controle. Para mais detalhes do método CUSUM, ver Veerkamp e Glas (2000) e Glas (2010).

Glas (2010) apresenta a estatística LM e o método CUSUM para analisar o ajuste do item e detectar mudanças ao longo do processo em CAT. LM foi testado para avaliar se o modelo da TRI da fase de pré-teste também se ajusta a fase on-line, bem como se a abordagem suporta a detecção de violações de modelos específicos.

O autor salienta que ambas as abordagens fornecem ferramentas práticas para monitoramento do DPI, que a estatística LM tem a vantagem de distribuições assintóticas conhecidas das estatísticas em que se baseia e que, por mais que estatísticas CUSUM tenham distribuições não

conhecidas, um valor crítico adequado pode ser encontrado via simulação, que é escolhido para fornecer a taxa de detecção desejada na situação prática, permitindo ajustar o procedimento às necessidades da situação específica.

Outros estudos sobre DPI em CAT podem ser obtidos em Abad et al. (2010), Bock, Muraki e Pfeiffenberger (1988), Guo e Wang (2003), McCoy (2010), Han e Guo (2011), Masters, Muckle e Bontempo (2009), Stocking (1988b), Zhu, Yu e Liu (2002), Zhang (2014), Zhang e Li (2016).

Em suma, após a seleção de um método ou um conjunto de métodos, as estimativas dos parâmetros dos itens devem ser monitoradas ao longo do tempo para determinar se houve mudanças quando comparado com administrações anteriores. Na prática, a decisão de utilizar um método particular, provavelmente levará em consideração a eficácia do método, a facilidade de implementação e a capacidade para comunicar os resultados às partes interessadas (CLARK, 2013).

#### *3.1.2.2.2 Possíveis causas do DPI e impactos*

Ao detectar DPI com o auxílio de testes estatísticos, sua causa deve ser investigada (VEERKAMP; GLAS, 2000; CLARK, 2013). O DPI pode ocorrer por várias razões, as quais estão relacionadas às mudanças culturais, educacionais e tecnológicas durante a vida útil de uma escala (BOCK; MURAKI; PFEIFFENBERGER, 1988). Essas mudanças podem ser sistemáticas ou não (CLARK, 2013), podendo, inclusive, estar relacionada aos fatores de *design* do teste (LI, 2008; MASTERS; MUCKLE; BONTEMPO, 2009).

A seguir, serão apresentadas algumas situações que podem ser responsáveis pelas mudanças nos parâmetros dos itens ao longo do tempo:

- Frequente exposição dos itens aos respondentes (VEERKAMP; GLAS, 2000; LI, 2008; MASTERS; MUCKLE; BONTEMPO, 2009; WALKER et al., 2010; CLARK, 2013; RISK, 2015), podendo levar ao pré-conhecimento do item;
- Falhas de segurança do teste - divulgação de itens pelos respondentes ou por outras formas de fraude, particularmente comum em situações de teste de alto impacto (HAN; GUO, 2011; CLARK, 2013; RISK, 2015), que também podem levar ao pré-conhecimento do item;
- Evento histórico (HAN; GUO, 2011);

- Divulgação nos meios de comunicação - pode influenciar o conhecimento geral sobre temas específicos, fazendo com que alguns itens apareçam menos exigentes com o tempo (CLARK, 2013);
- Mudanças no currículo (MASTERS; MUCKLE; BONTEMPO, 2009; GLAS, 2010; HAN; GUO, 2011; RISK, 2015); mudanças no construto ou conteúdo dos itens (CLARK, 2013; RISK, 2015) - mudanças nas áreas de conteúdo que ainda estão se desenvolvendo e mudando ano-a-ano ou quando itens já não medem adequadamente o construto de interesse, ou seja, estão desatualizados;
- Diferenças no modo de aplicação dos testes (P&P *x* CAT) (VEERKAMP; GLAS, 2000; MASTERS; MUCKLE; BONTEMPO, 2009; GLAS, 2010; CLARK, 2013);
- Mudanças na motivação dos respondentes (VEERKAMP; GLAS, 2000; GLAS, 2010);
- Mudanças no universo dos testes - público-alvo (participação por gênero, etnia e grupos de linguagem nativa) ou propósito da avaliação (WOLLACK; COHEN; WELLS, 2003; CLARK, 2013);
- Testes acelerados podem prejudicar a estimação dos parâmetros dos itens, afetar a qualidade da equalização e interpretação da escala (WOLLACK; COHEN; WELLS, 2003).
- Calibração inicial imprecisa dos itens pode resultar em DPI (RISK, 2015; HARMES; PARSHALL; KROMREY, 2003).

O DPI é considerado um desafio para o futuro dos testes (WALKER et al., 2010), pois quando ele existe, pode complicar o diagnóstico de habilidades específicas devido a itens que aparecem diferencialmente fácil ou difícil ao longo do tempo e, também, pode impactar na tomada de decisões em torno de um traço latente de corte ou quando discernir entre categorias de desempenho (CLARK, 2013).

A precisão dos parâmetros do item é muito importante em um CAT, pois cada aspecto do programa de teste é baseado nesses parâmetros, desde as funções de informação e a seleção de itens, até a estimativa do traço latente final. Se esses parâmetros são imprecisos ou instáveis, a integridade do CAT está em perigo (HARMES; PARSHALL; KROMREY, 2003).

O DPI pode ocorrer tanto em relação ao parâmetro de dificuldade quanto de discriminação do item. De acordo com Glas (2010), *drift* no parâmetro *c* raramente ocorre, pois na fase on-line os itens são adequados para o nível do traço latente dos respondentes. Bock, Muraki e Pfeiffenberger (1988) destacam que o DPI afetará as estimativas de dificuldade item a um grau mais forte do que as estimativas de discriminação item.

Wise e Kingsbury (2000) afirmam que, quando respondentes têm pré-conhecimento dos itens, ocorre um viés positivo na estimativa do traço latente; quando isso acontece para uma proporção substancial de respondentes, seus parâmetros acabam sendo inadequados para estimar o traço latente. Assim, quanto mais respondentes conhecem o conteúdo do item, o seu parâmetro de dificuldade torna-se mais fácil na escala, o parâmetro de discriminação se desloca em direção a zero e o parâmetro de acerto ao acaso se torna cada vez mais irrelevante (WISE; KINGSBURY, 2000; VEERKAMP; GLAS, 2000; LU; HAMBLETON, 2004).

Quando um item torna-se mais difícil de responder ao longo do tempo, pode causar um viés negativo na estimativa do traço latente. Isso pode ocorrer devido à mudanças educacionais, tecnológicas ou culturais (por exemplo, mudanças curriculares) (RISK, 2015).

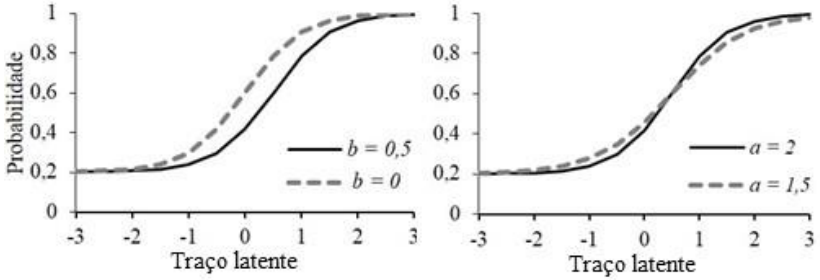
As Figuras 13 e 14 apresentam exemplos de CCI com DPI (linhas tracejada e pontilhada) em relação à sua CCI original com parâmetros $a = 2$, $b = 0{,}5$ e $c = 0{,}2$ (linha contínua). O *drift* foi de 0,5, tanto para *a* quanto para *b*, e *c* foi fixado em 0,2. Ao observá-las é possível perceber os diferentes impactos na probabilidade de resposta ao item para cada nível do traço latente dado os diferentes tipos de DPI.

A Figura 13 apresenta exemplos de CCI com *drift* apenas em um dos parâmetros para ML3P (*a* ou *b*). Já a Figura 14 apresenta *drift* nos parâmetros *a* e *b*, os quais são classificados como sendo na mesma direção ou em direções opostas. DPI na mesma direção ocorre quando parâmetros *a* e *b* diminuem (o item torna-se mais fácil e menos discriminativo). DPI em direções opostas ocorre quando *b* aumenta e *a* diminui (o item torna-se mais difícil e menos discriminativo) ou vice-versa.

Quando os parâmetros dos itens mudam, espera-se que o traço latente dos respondentes também sejam afetados (WELLS; SUBKOVIAK; SERLIN, 2002; CLARK, 2013). Estimativas do traço latente que são influenciadas pelo DPI podem causar sérios problemas não só ao nível individual do respondente, mas a nível do programa CAT

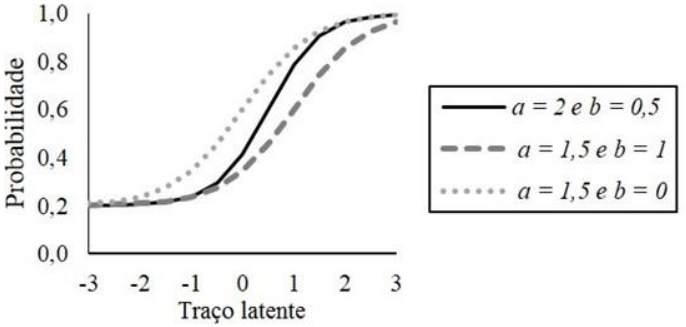
porque as estimativas dos traço latente são também utilizadas para calibrar novos itens; consequentemente, os itens do pré-teste que são calibrados com base nos traços latentes influenciados pelo DPI, podem não estar na mesma escala que outros itens já existentes no BI (HAN; GUO, 2011; GUO, 2016; LU; HAMBLETON, 2004).

Figura 13 – Exemplo de CCI com *drift* no parâmetro de dificuldade (à esquerda) e de discriminação (à direita).




Fonte: Elaborada pela autora.

Figura 14 – Exemplo de CCI com *drift* nos parâmetros de dificuldade e de discriminação do item.




Fonte: Elaborada pela autora.

De acordo com Han e Guo (2011), extensas pesquisas utilizando dados simulados e reais foram conduzidas para investigar o impacto do DPI nas estimativas dos parâmetros do item e do traço latente. No entanto, a maioria destes estudos de simulação assumem itens com DPI para todos os respondentes, o que não acontece no contexto de CAT, em que itens são expostos apenas a uma porcentagem de respondentes.

Para investigar este problema, os autores utilizaram simulação para analisar o impacto na calibração dos itens quando os itens com DPI foram expostos apenas para uma parte dos respondentes. Os resultados mostraram que o efeito a curto prazo do DPI na calibração de itens é limitada e inconsequente para as condições avaliadas (variando a quantidade de respondentes expostos aos itens com DPI), não influenciando significativamente nas estimativas do traço latente. Porém, poderia impactar ao longo do tempo.

*3.1.2.2.3 Tratamento aos itens com DPI*

Quando o DPI é detectado por meio dos métodos apresentados em 3.1.2.2.1 ou outros disponíveis, é preciso decidir o que fazer com estes itens. Para Wells et al. (2014), Veerkamp e Glas (2000), Walker et al. (2010), Thissen et al. (2007) e Clark (2013), deve-se identificar e excluir itens que causam impacto significativo na equalização e na estimação dos traços latentes. Preferencialmente, esta deve ser a ação tomada após a identificação em testes de alto impacto e é a ação destacada na Fase 2 - Etapa 2, principalmente se outros fatores também são identificados na Etapa 1.

Em um estudo feito por Risk (2015) que avaliou o impacto do DPI em diferentes condições de um teste de classificação adaptativo (aprovado/reprovado), os resultados indicaram consistentemente que, independentemente do tamanho do conjunto de itens ou do número de itens com DPI no BI, o *drift* de 1,0 *logit* teve o impacto mais negativo na precisão da medição. Portanto, a autora recomenda para as organizações de teste focalizarem seus recursos na identificação e correção de itens com grande magnitude de *drift* (acima de 1,0 *logit*).

Quando um item foi superexposto, por exemplo, é provável que mais respondentes irão respondê-lo corretamente do que na primeira vez que ele foi aplicado, esta diferença nos parâmetros indica que o item deve ser excluído (WALKER et al., 2010). Conforme Thissen et al. (2007), na área educacional, quando parâmetros dos itens sofrem alterações ao longo do tempo, necessitam ser substituídos.

A atualização regular dos itens é desejável tanto para manter o conteúdo frente às mudanças na educação e experiência da população de respondentes, quanto para proteger o teste de superexposição ou comprometimento por algum outro motivo (BOCK; MURAKI; PFEIFFENBERGER, 1988).

De acordo com Clark (2013), a decisão de manter um item com DPI no BI é aceitável quando se acredita que seja uma ocorrência temporária ou que o "desvio" representa uma mudança necessária para a escala, mas é preciso monitorá-los. Já Risk (2015) destaca que os profissionais podem optar por excluir o item totalmente do BI, recalibrar o item com a amostra atual de respondentes e mantê-lo no BI; ou podem retrabalhar o conteúdo do item ou opções de resposta e recalibrar o item como um item de pré-teste para os itens identificados.

# 4. PROCEDIMENTOS METODOLÓGICOS

Os procedimentos metodológicos estão subdivididos em três etapas: (1) revisão de literatura e suporte teórico para o desenvolvimento da sistemática; (2) especificação do BI real e definição do *design* do CAT operacional; e (3) aplicação do CAT e manutenção do BI.

## 4.1 REVISÃO DE LITERATURA E DESENVOLVIMENTO DA SISTEMÁTICA

Este trabalho caracteriza-se como teórico e constitui-se por um levantamento bibliográfico, em livros e artigos, para fundamentação do mesmo e para dar suporte ao desenvolvimento da sistemática. Para isso, diversas fontes foram consultadas.

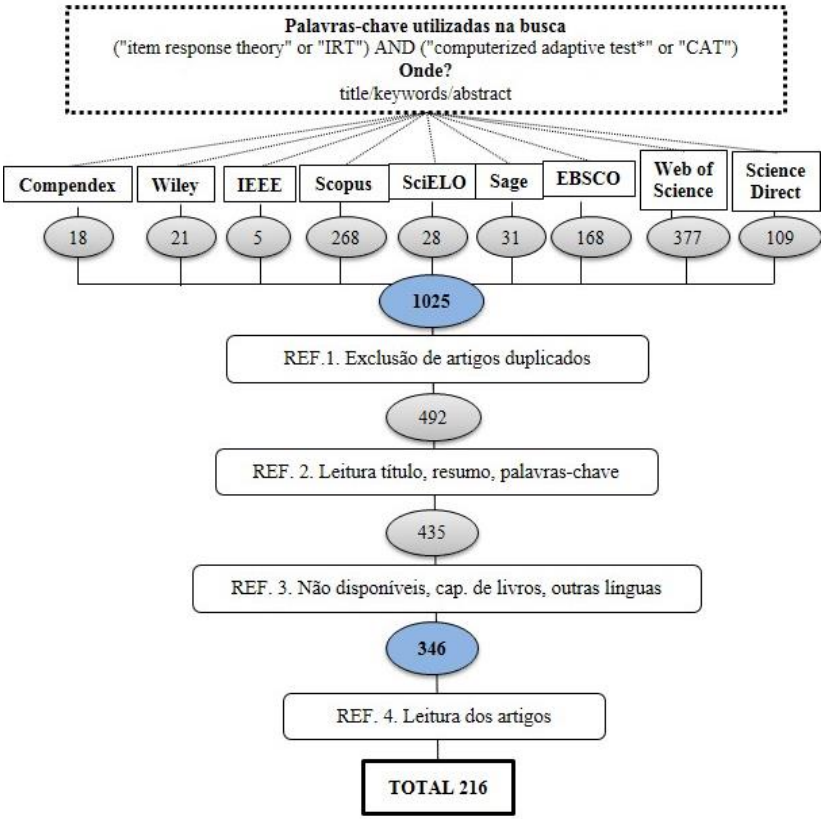
Primeiramente realizou-se uma busca de referencial abordando os temas CAT e TRI, em nove bases de dados, resultando na análise de 216 artigos, conforme mostra a Figura 15. Posteriormente, devido a falta de estudos para dar suporte a esta pesquisa, realizou-se uma segunda etapa de busca de referencial no *google* acadêmico, utilizando uma combinação de termos mais específicos como *computerized* (ou *computerised*) *adaptative testing* (ou *test*), *maintenance*, *maintaining*, *item pool* e *item bank*. Fez-se, também, uma análise das referências citadas nos trabalhos encontrados na literatura.

Por fim, o banco de referências (biblioteca) do *International Association of Computerized Adaptive Testing* (IACAT, 2015) foi consultado, o qual aborda trabalhos como capítulos de livros, artigos publicados em periódicos científicos e em eventos. Dentre os mais importantes, citam-se: *Annual Meeting of the National Council on Measurement in Education (NCME), Annual Meeting of the American Educational Research Association (AERA), Annual Conference of the International Association for Computerized Adaptive Testing (IACAT), Annual Meeting of the American Psychological Association (APA) e Annual Meeting of the Psychometric Society*. Desta forma, obteve-se uma extensa lista de referências sobre testes adaptativos computadorizados.

A partir da pesquisa bibliográfica, identificou-se as áreas que estão utilizando CATs e as vantagens obtidas por meio da aplicação desses testes, os componentes/regras de um CAT e os *softwares* atualmente utilizados no seu desenvolvimento e implantação. Também serviu de

suporte para justificar as etapas determinadas para a manutenção do BI ao longo do tempo e como operacioná-las, auxiliando na tomada de decisões.

Figura 15 – Estratégias e resultados das buscas realizadas nas bases de dados no período de início disponível pelas bases até agosto de 2015.




Fonte: Elaborada pela autora.

## 4.2 BI E DEFINIÇÃO DO *DESIGN* DO CAT

Esta seção detalha os procedimentos adotados para a definição do *design* do CAT e posterior aplicação dos testes em um curso de capacitação em Saúde Mental: álcool e outras drogas, oferecido na modalidade à distância pela Universidade Federal de Santa Catarina

(UFSC). Dados deste contexto foram utilizados com a intenção de apresentar uma situação real em que o CAT foi implementado e aplicado aos respondentes, necessitando de mantenção do BI.

Podem se inscrever neste curso profissionais que possuem nível médio, técnico ou superior completo e que atuam na Rede de Atenção Psicossocial (RAPS), tais como Centros de Convivência e Cultura, SAMU (Serviço de Atendimento Móvel de Urgência), Unidade de Pronto Atendimento (UPA) 24 horas, Pronto socorro em Hospital Geral, Unidade de Acolhimento, leitos de psiquiatria e saúde mental em Hospital Geral e Serviços Residenciais Terapêuticos, entre outros.

O teste é aplicado após a conclusão do curso, com aproximadamente 1.000 respondentes em cada edição de aplicação de testes, que acontece duas vezes por ano. O teste é disponibilizado no ambiente *Moodle* de ensino-aprendizagem, que deve ser acessado pelo cadastro do aluno.

Os seguintes *softwares* foram utilizados durante este trabalho: Bilog (ZIMOWSKI et al., 2003) para calibração dos itens; *software* R (R CORE TEAM, 2016) e o pacote catR (MAGIS; RAICHE, 2012) para as simulações relacionadas ao CAT; o Concerto (CONCERTO, 2016) para a aplicação do CAT aos respondentes, o qual também faz uso do *software* R e do pacote catR para definição do algoritmo de aplicação do teste. A dimensionalidade do BI foi verificada por meio da análise fatorial de informação completa (BOCK; GIBBONS; MURAKI, 1988) utilizando o pacote mirt (CHALMERS, 2012).

### 4.2.1 Composição do BI inicial e escala

O BI inicial consiste de 71 itens de múltipla escolha, com cinco alternativas de resposta; seus respectivos parâmetros, que foram calibrados pelo modelo ML3P da TRI para respostas dicotômicas (Equação 1); e do Módulo (conteúdo) ao qual cada item pertence. A escala desenvolvida tem por objetivo mensurar o conhecimento dos profissionais sobre os temas abordados durante o curso e que estão relacionados à Saúde Mental: Álcool e outras drogas.

Os itens foram desenvolvidos por especialistas da área e abordam sete Módulos: 1) Drogas e sociedade; 2) SUS, Políticas de Saúde Mental e Direitos Humanos; 3) Atenção Psicossocial e Cuidado; 4) Organização dos serviços para garantir acesso e promover vinculação do usuário de drogas; 5) Processo de trabalho no serviço de atenção à usuários de álcool

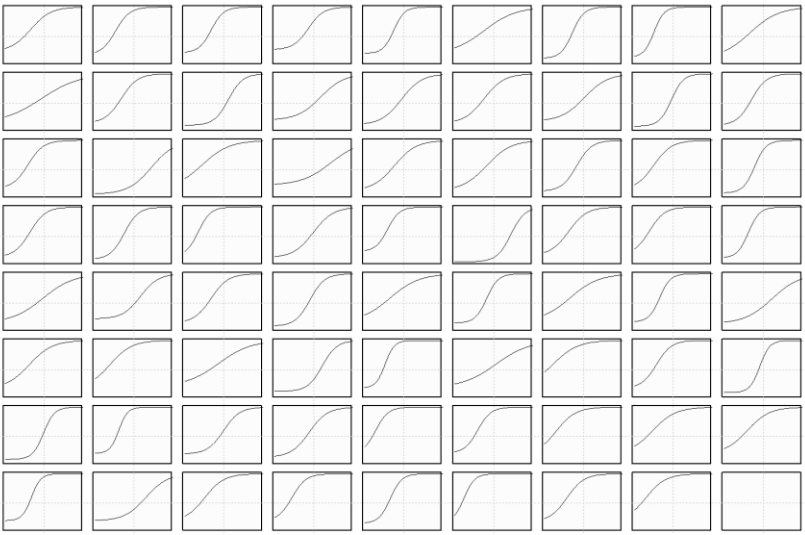
e outras drogas; 6) Recursos e Estratégias do Cuidado; e 7) Singularidades no Cuidado da Rede de Apoio Psicossocial.

O Módulo 1 é composto por 12 itens, Módulo 2 e Módulo 3 por 11 itens cada, Módulo 4 por 8 itens, Módulo 5 por 9 itens, Módulo 6 e Módulo 7 por 10 itens cada. A representação das CCIs dos 71 itens pode ser observada na Figura 16. De forma geral, os parâmetros de discriminação variaram de 0,67 a 2,95, com média de 1,62. Os parâmetros de dificuldade variaram entre -2,32 e 1,35, com média de -0,77. Os parâmetros de acerto ao acaso variaram de 0,07 a 0,28, com média de 0,18.

A escala de referência foi desenvolvida por meio de testes não-adaptativos, que foram aplicados via computador para profissionais de três turmas que concluíram o curso, no período de outubro de 2014 a março de 2015. Assim, utilizou-se a escala N(0,1), onde o zero da escala representa o conhecimento médio deste grupo de referência.

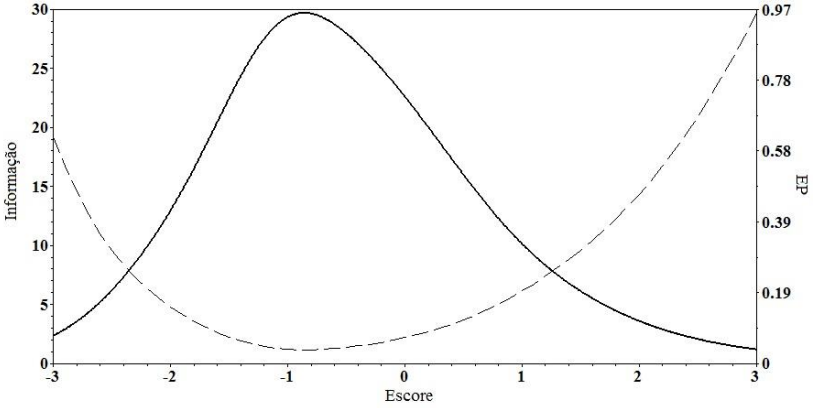
A Função de Informação do Banco de Itens (FIBI) inicial é apresentada na Figura 17. Observa-se que a curva se encontra deslocada à esquerda da média zero da escala de medida, ou seja, há mais itens fáceis do que itens difíceis no BI.

Figura 16 – CCIs dos 71 itens que compõem o BI inicial.



Fonte: Elaborada pela autora.

Figura 17 – FIBI com 71 itens.




Fonte: Elaborada pela autora.

A interpretação qualitativa da escala e a definição das habilidades por nível de conhecimento foi realizada por dois especialistas na área de Saúde Mental: Álcool e outras drogas, pelo método de definição de níveis âncora para posicionamento dos itens (ver ANDRADE; TAVARES; VALLE, 2000). A escala de conhecimento em Saúde Mental: álcool e outras drogas possui sete níveis âncora interpretáveis (-3, -2, -1, 0, 1, 2, 3), os quais foram classificados em níveis qualitativos de desempenho: insuficiente, básico, proficiente e avançado, conforme o Quadro 12.

### 4.2.2 Definição do *design* do CAT operacional

Diversos métodos foram comparados por meio de simulações com a finalidade de definir o *design* do CAT, antes de sua aplicação efetiva aos respondentes. Para tal, considerou-se um BI real, cujo traço latente que está sendo investigado é o conhecimento dos profissionais nos temas abordados durante o curso.

O teste, na sua concepção original, não é considerado de alto impacto. Por isso, a definição do *design* do CAT e a aplicação do teste não restringem os itens a uma taxa máxima de exposição. No entanto, simulações são efetuadas incluindo também esta restrição e seus impactos são avaliados. Por limitações relacionadas ao objetivo da avaliação e do BI, parte da sistemática de manutenção proposta neste trabalho foi

implementada durante a aplicação de duas edições de testes e os resultados serão analisados e apresentados.

Quadro 12 – *Feedback* geral dos níveis qualitativos da escala em Saúde Mental: Álcool e outras drogas.

| **Níveis qualitativos** | ***Feedback*** |
|---|---|
| **Insuficiente (abaixo de -3)** | Seu aprendizado sobre os conceitos básicos do campo do cuidado em saúde mental, álcool e outras drogas deve melhorar substancialmente. Sugerimos que retome as leituras dos capítulos e refaça as atividades. Sua participação é essencial para o avanço do cuidado em saúde mental, álcool e outras drogas. |
| **Básico (de -3 a 0)** | Você demonstrou domínio de conceitos básicos para atuar no campo da saúde mental, álcool e outras drogas que envolvem a organização dos processos de trabalho na RAPS, os fundamentos da clínica da atenção psicossocial e o desenvolvimento de atitudes e estratégias terapêuticas, bem como a necessidade de desconstruir estereótipos e realizar o acolhimento em situações de crise aos usuários de drogas. Precisa, agora, ampliar ainda mais seus estudos, suas reflexões e conhecimentos. Continue estudando, seu envolvimento com o cuidado das pessoas com uso problemático de álcool e outras drogas é muito importante! |
| **Proficiente (de 0 a 3)** | Você demonstrou ter adquirido os conhecimentos esperados pelo curso, para além dos conhecimentos básicos. Você mostrou-se capaz de compreender a atuação em equipe interdisciplinar, analisar a complexidade do campo da atenção psicossocial e as implicações das teorias e concepções aprendidas, planejar ações de prevenção, de redução de danos e posicionar-se frente aos estigmas e preconceitos associados aos usuários de drogas. Sugerimos que aprofunde seus estudos no sentido de qualificar ainda mais o cuidado em rede para as pessoas com uso problemático de álcool e outras drogas. Parabéns! |
| **Avançado (acima de 3)** | Você foi além da apreensão dos conhecimentos demandados pelo curso, Parabéns! Agora você pode contribuir para o avanço das políticas e do cuidado em saúde mental, álcool e outras drogas realizando estudos avaliativos e exercendo permanentemente a crítica sobre as práticas e teorias estabelecidas, a partir de pesquisas e produção de conhecimento. |

#### 4.2.2.1 Simulação de respondentes e matriz de respostas

Para avaliar a eficiência e precisão dos algoritmos que serão comparados sob diferentes regras e seus impactos na estimação do traço latente e no uso do BI, traços latentes verdadeiros para 1.000 respondentes foram gerados de uma distribuição normal, $\theta \sim N(0,1)$. Posteriormente, considerando o traço latente verdadeiro de cada respondente, as respostas aos 71 itens foram geradas a partir de uma distribuição de Bernoulli com parâmetros $p_{ij}$, para ML3P (Equação 1). Essa matriz de respostas foi armazenada para ser utilizada nos estudos simulados do CAT.

#### 4.2.2.2 Critérios para avaliação dos métodos

Buscando encontrar o melhor *design* para o CAT operacional, níveis de segurança do BI foram avaliados pelo comprimento médio do CAT, taxa de exposição dos itens, taxa de sobreposição de teste e uso dos itens do BI, que representa o número de itens não utilizados em cada condição.

Para avaliar os impactos na estimação dos traços latentes, utilizaram-se medidas de precisão como viés, RMSE, média do EP, variação do EP (EP máximo e mínimo), proporção de respondentes que atingiram a regra de parada do CAT com base na precisão preestabelecida e correlação entre os traços latentes verdadeiros e estimados.

O Viés e RMSE são calculados da seguinte maneira:

$$viés\left(\hat{\theta}\right) = \frac{1}{n}\sum_{j=1}^{n}\left(\hat{\theta}_j - \theta_j\right) \tag{19}$$

$$RMSE\left(\hat{\theta}\right) = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(\hat{\theta}_j - \theta_j\right)^2} \tag{20}$$

A taxa de exposição do item *i* ($TE_i$) é dada por:

$$TE_i = \frac{X_i}{n} \tag{21}$$

A taxa de sobreposição do teste (TS) é dada por:

$$TS = \frac{\sum_{i=1}^{I} (TE_i)^2}{\sum_{i=1}^{I} TE_i} \quad (22)$$

onde $n$ é número de respondentes; $\theta_j$ é o traço latente verdadeiro do respondente $j$ e $\hat{\theta}_j$ é a estimativa do traço latente para o respondente $j$; $X_i$ corresponde ao número de vezes que o item $i$ foi administrado entre todos os respondentes $j$.

A administração de todo o conjunto de itens aos respondentes serviu como uma condição de controle, resultando em uma matriz de dados completa para o CAT. Também, são fornecidas algumas representações gráficas e tabelas cujos resultados são condicionais por decil dos traços latentes. Ou seja, estatísticas condicionais da estimativa do traço latente em função dos decis dos verdadeiros níveis de traço latente. Ex.: viés médio condicional por decil.

#### 4.2.2.3 Estudos para definição do algoritmo

Nesta seção, apresenta-se como foram efetuados os seis estudos de simulação para definir o *design* do CAT operacional. Destaca-se que o resultado de um estudo auxilia na tomada de decisão do algoritmo testado nos estudos subsequentes. A restrição de BC sempre foi considerada nos algoritmos testados para o CAT, a não ser que seja especificado o contrário.

A restrição de BC é especificada pelo Módulo ao qual cada item pertence. Assim, itens são aplicados de forma proporcional a quatidade de itens de cada módulo no BI, conforme os seguintes passos (MAGIS; RAICHE, 2012):

1. Se nenhum item foi administrado, um subgrupo é escolhido aleatoriamente e o item ótimo é selecionado a partir deste subgrupo;
2. Se pelo menos um subgrupo ainda não foi alvo de seleção do item, um destes subgrupos é escolhido aleatoriamente e o item ótimo deste subgrupo é selecionado;

3. Se pelo menos um item por subgrupo já foi administrado, as proporções relativas empíricas dos itens administrados por subgrupo são computadas. Itens ótimos são selecionados dos subgrupos em que haja diferença entre a proporção empírica e o teórica até que elas sejam iguais. No caso de vários destes grupos, um grupo é escolhido aleatoriamente em primeiro lugar.

**ESTUDO 1: Condição de controle – aplicação de todos os itens do BI com variação do método de estimação do traço latente**

Este estudo visa identificar a precisão que pode ser obtida nos diferentes níveis do traço latente a partir do BI disponível para o CAT, bem como qual método produz estimativas mais próximas do verdadeiro traço latente quando o BI completo (71 itens) é aplicado aos respondentes. Assim, as seguintes regras foram definidas para as simulações do CAT:

- O teste inicia com um item mais informativo para $\theta = 0$ (item mediano da escala) e, em seguida, o traço latente é estimado para seleção do próximo item;
- Utilização do método de máxima informação de Fisher (MFI) para a seleção dos itens;
- Métodos testados para estimação do traço latente: EAP (com 19 pontos de quadratura), MAP, MV e WLE; Para os métodos bayesianos, uma distribuição *a priori* N(0,1) foi utilizada;
- Restrições: BC. Não foi estipulado nenhum método para controle da taxa de exposição dos itens;
- Regra de parada: comprimento fixo com 71 itens.

**ESTUDO 2: Definindo a regra de parada do CAT**

**1) Investigando a variação do EP nas estimativas do traço latente a cada item aplicado**

Buscou-se identificar o número de itens que precisava ser aplicado, de modo que o EP das estimativas do traço latente dos respondentes praticamente se estabilize. Para isso, utilizou-se como regra de parada o comprimento de 71 itens, MFI para seleção dos itens e EAP para estimação dos traços latentes. Assim, verificou-se em que momento (após quantos itens) a diferença/variação entre as precisões estimadas entre o item atual aplicado ($EP_{i,j}$) e o item anteriormente aplicado ($EP_{i-1,j}$) eram inferior aos critérios de 5%, 2,5%, 1,5% e 1% da precisão estimada do item anterior, conforme exemplo da equação a seguir:

$$\left(EP_{i,j} - EP_{i-1,j}\right) < 0{,}01 \times EP_{i-1,j} \quad (23)$$

onde $i$ corresponde ao item, $i$ = 1, ..., 71; e $j$ corresponde ao respondente, $j$ = 1,..., 1.000. $EP_{i,j}$ corresponde à estimativa do EP do traço latente do respondente $j$ após a aplicação de $i$ itens.

Uma pequena diferença na precisão indica que, mesmo aplicando mais itens, a precisão tende a se manter estável, provavelmente devido à falta de itens informativos restantes no BI para o nível do respondente. Assim, o teste pode ser encerrado, sem prejuízos à precisão do traço latente.

**2) CAT de comprimento variável**

Fez-se um estudo simulado considerando como regra de parada uma precisão mínima das estimativas dos traços latentes igual a 0,41 para dois comprimentos máximos do teste: "EP = 0,41 ou até 21 itens" e "EP = 0,41 ou até 35 itens". Essa precisão foi estabelecida com base na média do EP correspondente ao último decil (D10) das estimativas dos traços latentes quando o BI completo foi administrado em conjunto com o método EAP (Estudo 1).

Os comprimentos de 35 itens e de 21 itens foram testados porque foi o número de itens aplicados nos testes não-adaptativos do curso de capacitação. Além disso, forneceram bons resultados para a simulação anterior, que investigou a variação do EP com base nos diferentes critérios percentuais. Assim, para iniciar o teste, um item foi selecionado para $\theta = 0$ e os métodos MFI para seleção de itens e EAP para estimação do traço latente foram utilizados.

**ESTUDO 3: Testando diferentes itens para iniciar o CAT**

Neste estudo de simulação, considerou-se as seguintes regras para o CAT: método EAP para estimação do traço latente; método MFI para seleção dos itens; regra de parada com "EP = 0,41 ou até 21 itens". Foram testadas quatro regras para iniciar o CAT (*starts*), tais como:

- ***Start* 1:** sete itens iniciais fixos para todos os respondentes, sendo um item de cada Módulo, com bons parâmetros de discriminação ($a$ > 1,34) e dificuldade variando de -1,8 a 1,3;
- ***Start* 2:** cinco itens iniciais fixos com parâmetros de discriminação mais baixos (0,67 < $a$ < 0,91) e de dificuldade entre -1,03 e 1,32;

- ***Start* 3:** quatro itens iniciais para $\theta = 0$, podendo variar de -2 a 2. Neste caso, cada respondente pode receber diferentes itens para iniciar o teste;
- ***Start* 4:** três itens com bons parâmetros, $a > 1{,}1$ e $b = -1{,}1$; 0,08; 0,91 (fácil, mediano e difícil).

A finalidade deste estudo foi verificar os impactos nas estimativas e definir qual regra é mais adequada para ser implantada no CAT operacional. Para isso, diferentes parâmetros de dificuldade, de discriminação e número de itens iniciais aplicados antes da estimação do traço latente provisório foram testados. Esses itens iniciais pertenciam a Módulos distintos.

**ESTUDO 4: Definindo o método de seleção dos itens**

Quatro métodos de seleção de itens (MFI, bOpt, MPWI e MEPV – ver seção 2.2.3.2) foram analisados em conjunto com o método EAP de estimação do traço latente e regra de parada igual a "EP = 0,41 ou até 21 itens". O teste iniciou com *start* 4 (três itens iniciais fixos). A decisão por esses métodos de seleção de itens foi baseada nas características diferenciadas para selecionar os itens.

Considere $\theta = \hat{\theta}_{u_{i_1} \ldots u_{i_{k-1}}}$, $q$ o item de interesse (mas não previamente administrado), selecionado do subconjunto de itens pertencentes ao BI que podem ser apresentados ao respondente, com $i = 1, \ldots, I$ e $k = 1, \ldots, K$ é o *rank* dos itens em CAT, ou seja, $i_k$ é o índice do item administrado como o $k$-th item no teste para um respondente, $u_{i_q}$ corresponde à resposta. O método MFI seleciona o $k$-th item que maximiza a Equação 4 (seção 2.2.3.4), conforme abaixo (VAN DER LINDEN; PASHLEY, 2010):

$$i_k \equiv \arg\max_q \left\{ I_{U_q}\left( \hat{\theta}_{u_{i_1} \ldots u_{i_{k-1}}} \right) : q \in R_k \right\} \qquad (24)$$

O método *bOpt* seleciona o item seguinte com o nível de dificuldade igual à estimativa provisória do traço latente, obtida após a administração do item $k$.

Para o método MPWI, considere $I_{Uq}(\theta)$ a função de informação do item $q$ avaliado em $\theta$ e $L\left(\theta \mid u_{i_1}, \ldots, u_{i_{k-1}}\right)$ a função de verossimilhança

avaliada em $\theta$, dado o padrão de respostas aos $k$-1 itens administrados. Então, PWI para o item $q$ é dado por (MAGIS; RAICHE, 2012):

$$PWI_q = \int I_{U_q}(\theta) g(\theta) L(\theta \mid u_{i_1}, ..., u_{i_{k-1}}) d\theta \quad (25)$$

onde $g(\theta)$ é a distribuição *a priori* do traço latente.

Para MEPV, considere $P_q(\theta)$ a probabilidade de responder corretamente o item $q$ condicionado a $\theta$ e $Q_q(\theta) = 1 - P_q(\theta)$. A $Var(\theta \mid u_{i_1}, ..., u_{i_{k-1}}, 0)$ e $Var(\theta \mid u_{i_1}, ..., u_{i_{k-1}}, 1)$ são as variâncias posteriores de $\theta$, dado o padrão de respostas (atualizado pela resposta 0 e 1, respectivamente). Então, o EPV para o item $q$ é dado por (MAGIS; RAICHE, 2012):

$$EPV_q = P_q(\theta) Var(\theta \mid u_{i_1}, ..., u_{i_{k-1}}, 1) + Q_q(\theta) Var(\theta \mid u_{i_1}, ..., u_{i_{k-1}}, 0) \quad (26)$$

A variância é calculada como o erro padrão quadrático da estimativa EAP do traço latente, usando o padrão de respostas.

**ESTUDO 5: Comparando métodos de controle da taxa de exposição dos itens**

Este estudo visa verificar os impactos no uso dos itens do BI e nas estimativas do traço latente quando métodos de controle da exposição dos itens são inseridos. Conforme já mencionado, esta restrição não será implantada no CAT operacional do curso de capacitação. Porém, como é uma restrição muito importante para testes de alto impacto, seus impactos para este BI serão investigados.

Para tanto, os métodos elegibilidade do item - IE (VAN DER LINDEN; VELDKAMP, 2004) e restrito – MR (REVUELTA; PONSODA, 1998) foram comparados. Tais métodos são modificações na proposta de Sympson e Hetter (1985) e são baseados nas definições de dois eventos: (1) o item $i$ é selecionado pelo algoritmo CAT ($S_i$); e (2) o item $i$ é administrado ($A_i$). Como $A_i \subset S_i$, considera-se que $P(A_i) = P(A_i \mid S_i) P(S_i)$. As probabilidades $P(S_i)$ são determinadas pela composição do BI e a natureza do algoritmo CAT.

No entanto, a probabilidade condicional $P(A_i \mid S_i)$ são parâmetros de controle que devem ser configurados para $P(A_i) < r^{\max}$, sendo o $r^{max}$ um valor alvo determinado pela agência de teste (VAN DER LINDEN; VELDKAMP, 2004) [ver seção 2.2.3.3.3].

- **Método de elegibilidade de itens (IE)**

De acordo com Barrada, Abad e Veldkamp (2009), van der Linden e Veldkamp (2004), o método IE assemelha-se ao método SH na medida em que as decisões de impor restrições são probabilísticas. No entanto, o método não requer estudos de simulação demorados para estabelecer valores para os parâmetros de controle antes do uso operacional do teste, mas pode definir as probabilidades de inelegibilidade do item adaptativamente durante o teste usando as taxas reais de exposição do item (VAN DER LINDEN; VELDKAMP, 2004).

Tanto no método MR quanto em IE, os parâmetros $k_i$ são ajustados para cada novo respondente. Os parâmetros para o ($j$+1)-th respondente são calculados usando as taxas de exposição de quando o teste começa para o $j$-th respondente (BARRADA; ABAD; VELDKAMP, 2009):

$$k_i^{(j+1)} = \begin{cases} 1 & \text{se} \quad P^{(1\ldots j)}(A_i)/k_i^{(j)} \le r^{\max} \\ r^{\max} k_i^{(j)} / P^{(1\ldots j)}(A_i) & \text{se} \quad P^{(1\ldots j)}(A_i)/k_i^{(j)} > r^{\max} \end{cases} \qquad (27)$$

Neste método, os parâmetros $k_i$ são parâmetros $P(E_i)$, isto é, probabilidade do item $i$ ser elegível para o respondente. Enquanto no método SH os valores para os parâmetros $k_i$ pertencem ao intervalo [$r^{max}$, 1] e no método MR os valores possíveis são apenas {0, 1}, no método IE, $k_i$ pertence ao intervalo (0, 1].

Assim, antes de administrar qualquer item a um respondente é gerado um número aleatório pertencente ao intervalo uniforme (0, 1) para cada item e, somente se esse número for menor que o parâmetro $k_i$, esse item pertence ao subconjunto de itens elegíveis. Ao contrário do método MR, o método IE é de natureza probabilística; logo, a taxa de exposição máxima pode ser violada para alguns dos itens mais populares (BARRADA; ABAD; VELDKAMP, 2009).

- **Método de máxima informação restrita ou método restrito (MR)**

Conforme Revuelta e Ponsoda (1998), os itens são selecionados pelo método de máxima informação, mas nenhum deles pode ser exposto em mais de 100$k$% dos testes. Assim, supondo que um teste tenha sido administrado $t$ vezes e $A_i$ corresponde ao número de vezes que o item $i$ foi administrado nos $t$ testes anteriores, a taxa de exposição será $A_i$ /$t$.

O conjunto de itens disponíveis para o próximo teste será composto apenas de itens com taxas de exposição abaixo de $k$ e este conjunto de

itens disponíveis para administração muda de teste para teste. A única restrição para *k* é que, como alguns itens não poderão ser administrados em alguns testes, o valor de *k* deve ser maior do que o recíproco do quociente inteiro entre o tamanho do BI e o comprimento do teste (comprimento máximo do teste, em testes de comprimento variável), para garantir que haverá itens suficientes disponíveis para qualquer aplicação de teste (REVUELTA; PONSODA, 1998).

De acordo com Barrada, Abad e Veldkamp (2009), os parâmetros de controle podem adotar apenas dois valores, 0 e 1. O parâmetro $k_i$ será 0 se a taxa de exposição do item até o *j*-th respondente for maior ou igual a $r^{max}$, caso contrário, o parâmetro de controle será 1.

$$k_i^{(j+1)} = \begin{cases} 1 \text{ se } P^{(1\ldots j)}(A_i) < r^{\max} \\ 0 \text{ se } P^{(1\ldots j)}(A_i) \geq r^{\max} \end{cases} \quad (28)$$

Assim, MR adapta o subconjunto do BI que está disponível para administração para cada respondente. Os parâmetros $k_i$ são utilizados para definir o subconjunto de itens disponíveis para administração (assim, como os parâmetros $P(E_i)$). Neste método, as taxas de exposição dos itens não ultrapassam $r^{max}$, mesmo quando há mudanças na composição do BI ou na distribuição do traço latente (BARRADA; ABAD; VELDKAMP, 2009).

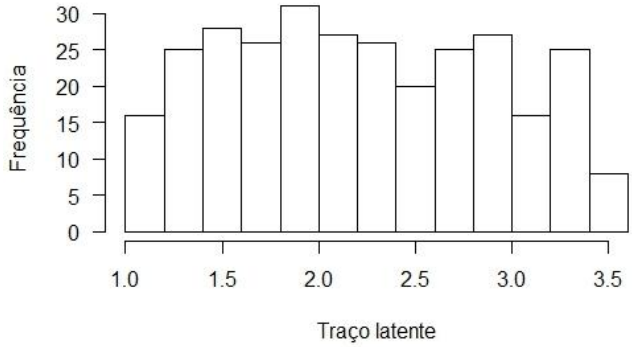
Em conjunto com os métodos de controle da taxa de exposição, utilizou-se o método EAP para estimar o traços latente e MFI para a seleção de itens; o teste iniciou com *start* 4 (três itens iniciais) e a regra de parada foi "EP=0,41 ou até 21 itens".

Devido à restrição para o parâmetro *k* em MR, ou seja, *k* deve ser maior do que o recíproco do quociente inteiro entre o tamanho do BI e o comprimento do teste [logo, $k > 0{,}3$ (21/71)], optou-se por fixar uma taxa de exposição máxima igual a 0,4, pois o BI é pequeno. Assim, o item deve ser apresentado, no máximo, para 40% dos respondentes.

**ESTUDO 6: CAT para respondentes com elevado nível do traço latente**

Considerando as peculiaridades do BI utilizado neste trabalho, fez-se uma nova simulação de 300 traços latentes verdadeiros, porém, obtendo apenas valores elevados, ou seja, eles foram gerados de uma distribuição uniforme, variando de 1 a 3,5, conforme mostra a Figura 18.

Figura 18 – Histograma dos traços latentes verdadeiros considerando apenas níveis elevados.




Fonte: Elaborada pela autora.

Este estudo teve por objetivo investigar o impacto nas estimativas dos traços latentes quando o BI não possui muita informação no extremo superior da escala, bem como o uso dos itens do BI, sob diferentes regras que incluem BC e controle da taxa de exposição.

Para tanto, considerou-se as seguintes regras do CAT: três itens fixos para iniciar o teste (*start* 4); método MFI para seleção dos itens; e método EAP para estimação do traço latente dos respondentes. As seguintes situações foram comparadas:

**(1)** o BI completo foi aplicado;

**(2)** CAT com regra de parada baseada no "EP=0,41 ou até 21 itens", com BC e sem restrição da taxa de exposição dos itens;

**(3)** CAT com regra de parada baseada no "EP=0,41 ou até 21 itens", com BC e com restrição da taxa de exposição dos itens em 0,40 (métodos MR e IE);

**(4)** CAT com regra de parada baseada no "EP=0,41 ou até 21 itens", sem BC e sem restrição da taxa de exposição dos itens.

## 4.3 APLICAÇÃO DO CAT E MANUTENÇÃO DO BI

Uma vez desenvolvido o BI e definido o *design* do CAT operacional com base nas simulações efetuadas nos estudos anteriores, o CAT foi implementado e aplicado em duas edições de aplicação de testes: dezembro de 2015 e junho de 2016. O número esperado de respondentes

em cada edição não era muito grande (de 700 a 1.000 respondentes) e o BI possuía 71 itens.

Os principais resultados dessas aplicações dos CATs são apresentados em conjunto com a descrição da implementação de algumas etapas da sistemática para a manutenção do BI (conforme seção 3.1). As regras implementadas na primeira edição de aplicação dos CATs operacionais foram:

- Três itens iniciais fixos com bons parâmetros, $a > 1{,}1$ e $b = -1{,}1$; 0,08 e 0,91 (*Start* 4);
- Método MFI para seleção de itens;
- Método EAP para estimação dos traços latentes provisório e final, com distribuição *a priori* N(0,1) e 19 pontos de quadratura;
- Restrição de balanceamento de conteúdo (sete Módulos).
- Regra de parada: "EP = 0,41 ou até 21 itens".

O algoritmo foi o mesmo para as duas edições de testes; exceto para os três itens iniciais, que foram substituídos por outros três itens com o mesmo padrão de dificuldade, os quais pertenciam a Módulos distintos.

O teste adaptativo estava disponível para os respondentes via *Moodle*, cujo acesso era por meio do *login* de cada profissional que participou do curso de capacitação. O teste ficou "no ar" durante um período determinado (em torno de um mês) e os profissionais podiam responder ao teste de suas casas ou de outro local, bastando estar conectados à internet. Não foi inserida a restrição de tempo máximo predefinido para completar o teste, nem para responder aos itens.

O respondente estava impossibilitado de omitir respostas e passar para o próximo item. Essa ação poderia expor os itens sem necessidade e, também, poderia levar a ação por parte do respondente de ignorar vários itens até que um de seu conhecimento aparecesse na tela. Por isso, caso o indivíduo tentasse prosseguir no teste sem responder, uma mensagem aparecia na tela avisando que a resposta seria considerada como incorreta.

Para receber o traço latente pela TRI e as interpretações do seu nível de proficiência, o respondente precisava encerrar o teste. Caso ele resolvesse interromper o teste e posteriormente retornar, ele daria continuidade do local onde parou. Além disso, como o indivíduo poderia responder o teste mais de uma vez para testar seus conhecimentos, quando ocorrido, apenas os dados da primeira tentativa foram considerados para as análises.

## 5. ANÁLISE DOS RESULTADOS

Os resultados são apresentados em duas etapas: (1) resultados dos estudos simulados para definição do *design* do CAT operacional; e (2) resultados da aplicação do CAT em conjunto com algumas etapas de manutenção do BI, em duas edições de testes adaptativos. Em todos os estudos, a restrição de balanceamento de conteúdo foi considerada, exceto quando especificado o contrário.

### 5.1 DEFINIÇÃO DO DESIGN DO CAT

**ESTUDO 1: Condição de controle – aplicação de todos os itens do BI com variação do método de estimação do traço latente**

A Tabela 1 apresenta os principais resultados para os quatro métodos de estimação que foram investigados: EAP, MAP, MV e WLE, quando o BI completo foi aplicado a todos os respondentes. Pode-se observar que os métodos bayesianos produziram resultados bem diferentes dos métodos clássicos.

Dentre os métodos bayesianos, EAP produziu menor RMSE e menor viés. Dentre os métodos clássicos, tem-se que WLE foi um pouco melhor do que MV nas estatísticas analisadas.

Tabela 1 – Comparação dos métodos de estimação do traço latente para o BI completo.

| Método | Correlação | RMSE | Viés médio | EP médio | EP máximo | EP mínimo |
|---|---|---|---|---|---|---|
| EAP | 0,964 | 0,272 | -0,021 | 0,261 | 0,539 | 0,145 |
| MAP | 0,964 | 0,276 | -0,035 | 0,258 | 0,538 | 0,190 |
| MV | 0,954 | 0,330 | 0,018 | 0,294 | 1,660 | 0,193 |
| WLE | 0,957 | 0,303 | -0,017 | 0,283 | 1,379 | 0,193 |

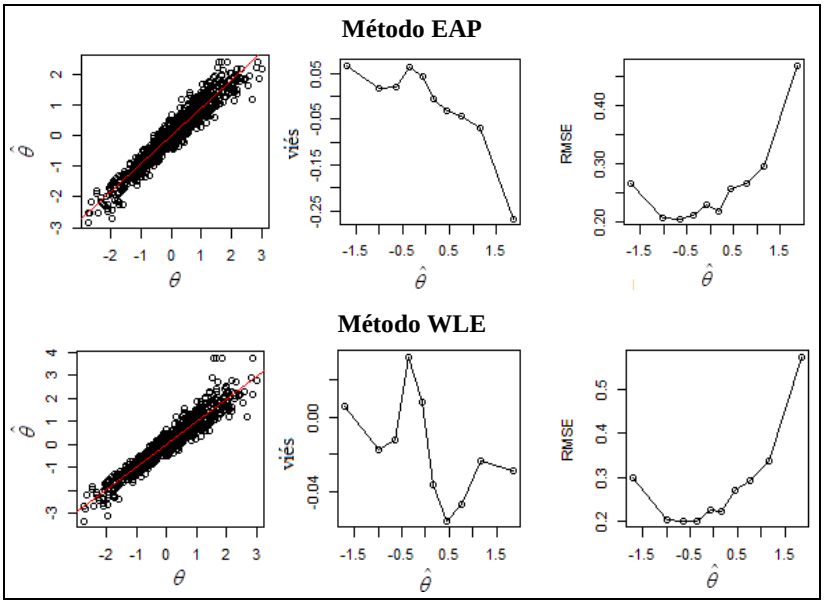
Fonte: Elaborada pela autora.

Comparando os métodos clássicos com os métodos bayesianos na Tabela 1, nota-se que os métodos clássicos produziram o menor viés e o método MV foi o único a gerar um viés médio positivo, ou seja, superestimou as estimativas. Para as demais estatísticas investigadas, os métodos bayesianos foram superiores na recuperação dos parâmetros verdadeiros dos respondentes e geraram uma correlação maior entre os traços latentes estimados e verdadeiros.

Além disso, o EP máximo para os métodos clássicos foram significativamente maiores do que nos bayesianos. Essa discrepância se torna evidente para os níveis com pouca informação na escala, ou seja, acima de 1,4. Esses indivíduos acertaram todos os itens ou quase todos; logo, os métodos clássicos apresentam problemas de estimação nesses casos e métodos bayesianos são preferíveis (VAN DER LINDEN; PASHLEY, 2010).

A Figura 19 apresenta o gráfico de dispersão entre os valores estimados $(\hat{\theta})$ e verdadeiros dos traços latentes $(\theta)$ para os métodos EAP e WLE, os quais produziram melhores estimativas. Observa-se, também, o viés e o RMSE condicionais dos decis dos traços latentes para ambos os métodos. É possível notar que o método EAP produz as melhores estimativas no extremo superior da escala; o RMSE dos métodos EAP e WLE são parecidos, mas o comportamento do viés é bastante diferente entre os métodos, mostrando os intervalos em que há superestimação ou subestimação dos verdadeiros traços latentes em cada método.

Figura 19 – Resultados condicionais das estimativas dos traços latentes para os métodos EAP e WLE com BI completo.




Fonte: Elaborada pela autora.

Com base nesses resultados, constatou-se que o método EAP é mais eficiente e deve ser utilizado no CAT operacional para estimação dos traços latentes dos respondentes. Assim, a Tabela 2 apresenta os resultados detalhados para este método com o BI completo, os quais são separados em decis. Cada decil é formado por dados referentes aos traços latentes de 100 respondentes.

Conforme esperado, o maior EP médio é obtido para D10 (EP = 0,414), que é composto por estimativas de traços latentes acima de 1,3. Também é possível notar que o viés médio dos decis são positivos até D5; após, quando os traços latentes estão acima da média zero, o viés médio passa a ser negativo, ou seja, os traços latentes são subestimados, uma vez que há poucos itens no BI para esta região.

Tabela 2 – Resultados das estimativas dos traços latentes separados em decis para o método EAP com BI completo.

| **Estatísticas** | **Decil** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **D1** | **D2** | **D3** | **D4** | **D5** | **D6** | **D7** | **D8** | **D9** | **D10** |
| Traço latente médio | -1,695 | -1,002 | -0,636 | -0,349 | -0,075 | 0,167 | 0,435 | 0,769 | 1,153 | 1,853 |
| RMSE | 0,266 | 0,206 | 0,205 | 0,211 | 0,228 | 0,219 | 0,256 | 0,266 | 0,295 | 0,469 |
| Viés médio | 0,066 | 0,016 | 0,020 | 0,064 | 0,043 | -0,005 | -0,032 | -0,044 | -0,07 | -0,267 |
| EP médio | 0,242 | 0,190 | 0,198 | 0,207 | 0,224 | 0,239 | 0,262 | 0,296 | 0,341 | 0,414 |

Fonte: Elaborada pela autora.

**ESTUDO 2: Definindo a regra de parada do CAT**

Investigou-se a variação do EP das estimativas dos traços latentes a cada item apresentado ao respondente, sob quatro critérios. Tais critérios auxiliaram na definição do comprimento máximo do CAT operacional. Os seguintes resultados gerais foram obtidos:

- **Critério de 5%:** Este critério permite uma variação maior entre os EPs das estimativas entre os itens aplicados, sendo o critério menos rigoroso. Por isso, do 9º para o 10º item, a diferença entre as precisões foi de 5% do valor com 9 itens (Equação 23). Ou seja, com um comprimento de 10 itens, todos os respondentes atingiram o critério de 5% de variação entre EPs.
- **Critério de 2,5%:** todos os respondentes atingiram este critério com até 17 itens;
- **Critério de 1,5%:** todos os respondentes atingiram este critério com até 23 itens, sendo que apenas quatro deles (0,4%)

precisaram responder mais do que 21 itens para atingir este critério;

- **Critério de 1%:** Este é o critério mais rigoroso e exige que as estimativas estejam mais estáveis. Logo, todos os respondentes atingiram este critério com até 34 itens, sendo que apenas 60 respondentes (6%) precisaram responder acima de 21 itens.

Considerando os critérios mais cautelosos de 1,5% e 1%, pode-se avaliar a possibilidade de um CAT com comprimento de 35 itens ou 21 itens, pois mostraram bons resultados nesta etapa e, também, foram os comprimentos utilizados nos testes não-adaptativos para o desenvolvimento do BI.

Como a ideia é fazer uso dos benefícios do CAT para fornecer testes mais curtos e, pensando em testes de alto impacto, para não expor desnecessariamente os itens, essas opções de comprimento máximo foram investigadas em conjunto com a obtenção de um EP = 0,41 nas estimativas dos traços latentes como regra de parada do CAT.

Este EP de estimação corresponde às estimativas de traços latentes via EAP do último decil (Tabela 2) quando o BI completo foi aplicado. Este EP não é muito pequeno, mas, se um EP muito baixo fosse estipulado, muitos respondentes nos extremos da escala não atenderiam o critério devido a qualidado no BI real utilizado e muitos itens seriam expostos desnecessariamente.

A Tabela 3 apresenta os resultados comparativos de diversas estatísticas considerando, então, um CAT de comprimento variável, cuja regra de parada é "EP=0,41 ou até 35 itens" ou "EP = 0,41 ou até 21 itens". Observa-se que os resultados são semelhantes para a maioria das estatísticas investigadas. Isso indica que, mesmo tendo respondido até 14 itens a mais em um dos casos e a proporção de respondentes que atingiram a precisão predefinida de 0,41 tenha sido diferente (0,919 para até 35 itens e 0,825 para até 21 itens), não obteve-se diferença significativa nas demais estatísticas relacionadas à precisão, o que também pode ser confirmado pela variação do EP.

A partir desses resultados, nota-se que, quando o BI completo foi apresentado aos respondentes, o EP máximo obtido para o método EAP foi de 0,539 (Tabela 1), estando próximo ao obtido pela regra de parada "EP=0,41 ou até 21 itens", igual a 0,57. Para esta regra, 185 indivíduos responderam aos 21 itens, sendo que 175 deles não

atingiram a regra de parada EP=0,41, que corresponde a 17,5% dos respondentes. Além disso, 11 itens do BI não foram utilizados.

Tabela 3 – Resultados gerais da comparação do número máximo de itens quando a regra de parada é baseada na precisão (EP=0,41).

| **Estatísticas** | **Comprimento máximo do CAT** | |
|---|---|---|
| | **35 itens** | **21 itens** |
| Comprimento médio do CAT | 14,13 | 12,58 |
| Nº mínimo de itens aplicados | 7 | 6 |
| Correlação | 0,916 | 0,916 |
| RMSE | 0,408 | 0,408 |
| Viés | -0,039 | -0,034 |
| Proporção de respondentes que atingiram regra de parada | 0,919 | 0,825 |
| Variação do EP | 0,36 a 0,55 | 0,36 a 0,57 |
| Taxa de sobreposição | 0,438 | 0,456 |

Fonte: Elaborada pela autora.

Para a regra de parada do CAT definida por "EP=0,41 ou até 35 itens", 82 indivíduos responderam aos 35 itens, sendo que 81 deles não atingiram a regra de parada EP=0,41, que corresponde a 8,1% dos respondentes. Neste caso, dois itens não foram utilizados. Por fim, com base nesses estudos, optou-se por utilizar como regra de parada no CAT operacional, "EP=0,41 ou até 21 itens".

**ESTUDO 3: Testando diferentes itens para iniciar o CAT**

A Tabela 4 apresenta os resultados de diferentes combinações de itens para iniciar o CAT, considerando os métodos EAP de estimação, MFI para seleção de itens, regra de parada igual a "EP=0,41 ou até 21 itens" e restrição de BC.

Observa-se que *Start* 4 apresentou os melhores resultados em relação ao RMSE, viés e correlação com o traço latente verdadeiro, do que as outras formas testadas para iniciar o CAT. *Start* 4 contém três itens iniciais com $b$ = -1,1; 0,08; 0,91 e possuem bons parâmetros de discriminação ($a$ > 1,1).

Em relação ao uso dos itens do BI, o *start* 4 apresentou o 2º melhor desempenho, ficando atrás do *Start* 3, que insere aleatoriedade na seleção dos itens iniciais e, consequentemente, apresentou a menor taxa de sobreposição de itens e fez melhor uso dos itens disponíveis no BI. Por

outro lado, também apresentou o pior RMSE e viés dentre os métodos avaliados. Devido a esses resultados, os itens de *start* 4 foram utilizados no algoritmo do CAT operacional.

*Start* 4 sem BC gerou o menor uso dos itens do BI, com 20 itens que não foram selecionados para aplicação (aproximadamente 28% do BI). Assim, itens mais informativos para os respondentes foram selecionados, fazendo com que a proporção de indivíduos que atingiram a regra de parada EP=0,41 fosse a maior dentre os métodos (85,7%). Por outro lado, não houve melhora nos valores de RMSE, viés e taxa de sobreposição, se comparado ao *start* 4 (com BC).

**ESTUDO 4: Definindo o método de seleção dos itens**

Neste estudo, cinco métodos de seleção de itens foram comparados (MFI, bOpt, MPWI e MEPV) em conjunto com *start* 4 para iniciar o CAT, método EAP de estimação dos traços latentes e regra de parada "EP=0,41 ou até 21 itens". Os resultados gerais são apresentados na Tabela 5. De forma geral, tem-se:

- Os métodos de seleção de itens MPWI e MEPV tiveram um tempo de simulação (em minutos) muito superior aos demais. Isso pode fazer com que o tempo entre o envio da resposta e o aparecimento do próximo item na tela seja demasiado, tornando-se um problema nas avaliações.
- bOpt apresentou um comprimento médio de teste maior do que os outros métodos e todos os itens do BI foram expostos pelo menos uma vez aos respondentes. Também apresentou os piores valores em relação ao RMSE, correlação, taxa de sobreposição de itens e proporção de respondentes que atingiram a regra de parada com base na precisão do traço latente.
- MPWI apresentou os melhores resultados. No entanto, seu tempo é elevado. MFI e MPWI apresentaram os menores valores para o viés.

A partir desses resultados, optou-se por continuar utilizando o método EAP em conjunto com MFI para a seleção de itens, dado que este método apresentou a menor taxa de sobreposição de itens e, também, apresentou valores próximos aos métodos MPWI e MEPV; já bOpt apresentou os piores resultados. Assim, esta regra foi implantada no CAT operacional.

Tabela 4 – Resultados gerais da comparação de diferentes combinações para iniciar o CAT.

| ***Start*** | **comp. médio** | **Nº mín. itens** | **Cor.** | **RMSE** | **Viés** | **Variação do EP** | | **Prop. de indivíduos que atingiram EP=0,41** | **Uso do BI** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | **Mín.** | **Máx.** | | **Taxa sobrep.** | **Nº de itens não usados** | **Nº de itens expostos a todos** |
| **1** (7 itens) | 15,1 | 10 | 0,911 | 0,418 | -0,021 | 0,357 | 0,583 | 0,796 | 0,659 | 14 | 7 |
| **2** (5 itens) | 15,66 | 11 | 0,913 | 0,416 | -0,022 | 0,365 | 0,580 | 0,796 | 0,625 | 12 | 5 |
| **3** (4 itens) | 14,07 | 7 | 0,902 | 0,439 | -0,035 | 0,353 | 0,584 | 0,812 | 0,346 | 0 | 0 |
| **4** (3 itens) | 14,36 | 9 | 0,914 | 0,413 | -0,020 | 0,347 | 0,575 | 0,814 | 0,562 | 11 | 3 |
| **4** (sem BC) | 12,54 | 8 | 0,911 | 0,420 | -0,026 | 0,363 | 0,569 | 0,857 | 0,575 | 20 | 3 |

Fonte: Elaborada pela autora.

Tabela 5 – Resultados gerais da comparação de diferentes métodos de seleção de itens.

| **Métodos** | **comp. médio** | **Nº mín. itens** | **Cor.** | **RMSE** | **Viés** | **Variação do EP** | | **Prop. de indivíduos que atingiram EP=0,41** | **Tempo (min.)** | **Taxa sobrep.** | **Nº itens não usados** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | **Mín.** | **Máx.** | | | | |
| **MFI** | 14,36 | 9 | 0,912 | 0,417 | -0,019 | 0,351 | 0,574 | 0,810 | 2,29 | 0,561 | 11 |
| **bOpt** | 17,21 | 10 | 0,904 | 0,435 | -0,021 | 0,370 | 0,577 | 0,747 | 2,67 | 0,598 | 0 |
| **MPWI** | 14,38 | 9 | 0,915 | 0,409 | -0,019 | 0,350 | 0,573 | 0,814 | 10,78 | 0,587 | 14 |
| **MEPV** | 14,31 | 9 | 0,915 | 0,411 | -0,022 | 0,346 | 0,573 | 0,811 | 30,87 | 0,576 | 14 |

Fonte: Elaborada pela autora.

**ESTUDO 5: Comparando métodos de controle da taxa de exposição dos itens**

A Tabela 6 apresenta os resultados gerais da comparação dos métodos IE e MR para controle da exposição dos itens, limitando-os a taxa máxima de 0,4. Os resultados para os métodos foram muito semelhantes para todas as estatísticas analisadas.

Tabela 6 – Comparação dos métodos MR e IE para controle da taxa de exposição dos itens.

| Estatísticas | Métodos | |
|---|---|---|
| | **MR** | **IE** |
| comp. médio | 15,36 | 15,40 |
| Nº mín. itens | 9 | 9 |
| Cor. | 0,910 | 0,908 |
| RMSE | 0,421 | 0,426 |
| Viés | -0,023 | -0,025 |
| Variação do EP | 0,347 a 0,630 | 0,345 a 0,614 |
| Prop. de indivíduos que atingiram EP=0,41 | 0,729 | 0,733 |
| Tempo (min.) | 2,02 | 2,07 |
| Taxa de sobreposição | 0,452 | 0,455 |
| Nº itens não usados | 3 | 3 |

Fonte: Elaborada pela autora.

MR apresentou valores levemente melhores para as seguintes estatísticas: correlação, RMSE, viés, tempo de simulação e taxa de sobreposição. O método IE apresentou melhores valores para o EP máximo obtido (0,614) e para a proporção de respondentes que atingiram a regra de parada EP=0,41.

Nos dois casos, apenas três itens do BI não foram utilizados e o comprimento mínimo de teste foi de nove itens. Em MR, seis itens atingiram a taxa máxima de exposição igual a 0,4; já no método IE, foram três itens. Nos dois métodos, esses itens que atingiram maior taxa de exposição possuíam níveis de dificuldades variados, mas principalmente entre -1 e 0. Os três itens não utilizados foram os mesmos para os dois métodos.

Comparando os resultados da Tabela 6 com os resultados do Método MFI sem restrição da taxa de exposição dos itens (Tabela 5), observa-se que a correlação é maior, RMSE, viés, EP máximo,

proporção de respondentes que atingiram a regra de parada são melhores e o comprimento médio do teste é menor na Tabela 5.

No entanto, inserindo esta restrição para a exposição dos itens, a taxa de sobreposição de itens cai de 0,561 para 0,45, assim como o uso dos itens do BI é melhor, ou seja, o número de itens não utilizados é menor. Por isso, conforme destacado na literatura, o preço para se ter maior segurança dos itens é a redução na precisão da medida.

Esses resultados também podem ser observados na Figura 20, por meio do gráfico de dispersão entre os traços latentes verdadeiros e estimados e a distribuição da taxa de exposição dos itens em relação ao parâmetro de discriminação do item para MR, IE e sem controle da taxa de exposição. Conforme esperado pela definição dos métodos, MR conseguiu controlar a taxa máxima de exposição dos itens em 0,40; já no método IE, os itens utilizados no CAT apresentaram uma taxa de exposição um pouco acima de 0,40.

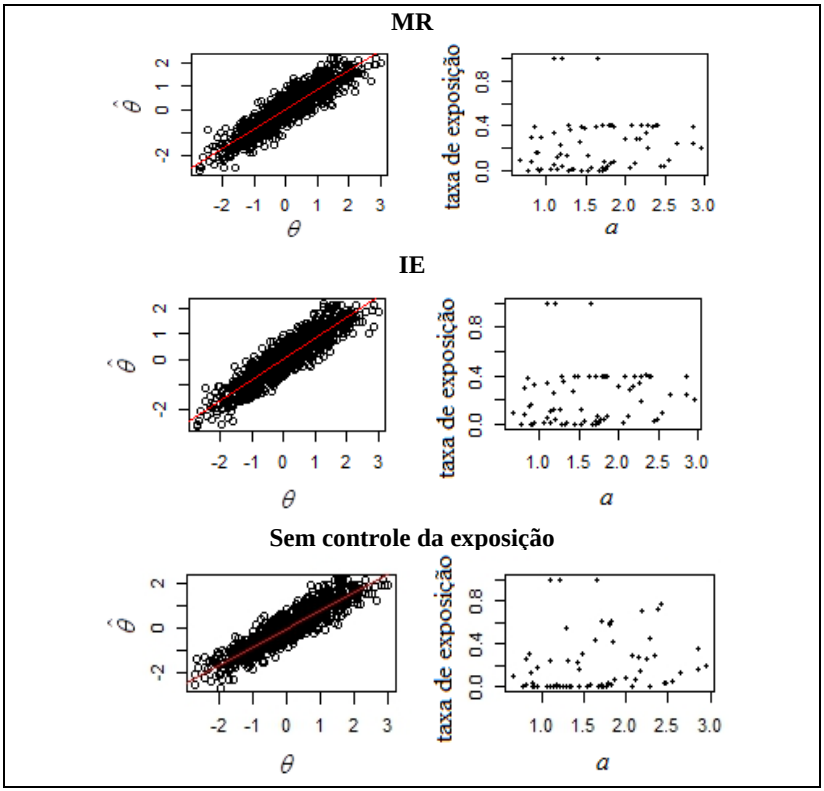
Os três itens que aparecem com taxa máxima de exposição, na Figura 20, são os três itens iniciais fixos para todos os respondentes. Quando não é inserido o controle da exposição, itens mais discriminativos são mais expostos. Também, o número de itens não utilizados cai de 11 para 3 itens quando a taxa máxima de exposição é fixada.

Em testes de alto impacto, taxas máximas mais utilizadas na literatura variam de 15% a 30% (ver ABAD et al., 2010; STOCKING, 1994; ALI, CHANG, 2014; BARRADA et al., 2009). No entanto, neste trabalho, não seria possível inserir uma taxa máxima desta magnitude, uma vez que o BI é pequeno.

**ESTUDO 6: CAT para respondentes com elevados níveis do traço latente**

Considerou-se apenas respondentes com traço latente verdadeiro acima de 1 para investigação dos impactos no uso dos itens do BI e na estimação dos traços latentes. As seguintes regras foram testadas: **(1)** o BI completo foi aplicado; **(2)** CAT com regra de parada baseada no "EP=0,41 ou até 21 itens", com BC e sem restrição da taxa de exposição; **(3)** CAT com regra de parada baseada no "EP=0,41 ou até 21 itens", com BC e com restrição da taxa de exposição dos itens em 0,40 (métodos MR e IE); **(4)** CAT com regra de parada baseada no "EP=0,41 ou até 21 itens", sem BC e sem restrição da taxa de exposição dos itens. Os resultados são apresentados na Tabela 7.

Figura 20 – Inserindo controle da exposição dos itens: acurácia nas estimativas e distribuição da taxa de exposição de acordo com o parâmetro de discriminação.




Fonte: Elaborada pela autora.

Os resultados obtidos para as regras do CAT sem a restrição da taxa de exposição dos itens (regras 1, 2 e 4 na Tabela 7), foram próximos, em magnitude, para a maioria das estatísticas analisadas. Nesses cenários, nota-se que o viés médio das estimações foi significativamente diferenciado (0,118, -0,112 e -0,026).

Conforme esperado, os melhores resultados foram obtidos para a regra do CAT sem BC, baseado na regra de parada um "EP=0,41 ou até 21 itens" (regra 4 na Tabela 7). No entanto, este apresentou uma elevada taxa de sobreposição de itens (0,981); isso também ocorre quando há BC (0,977).

Tabela 7 – Resultados da comparação de algoritmos para respondentes com traço latente elevado.

| | Regra | Cor. | RMSE | Viés | Taxa sobrep. | Nº itens com máxima expos. | Nº de itens não utilizados | EP médio | EP máx. | EP mín. |
|---|---|---|---|---|---|---|---|---|---|---|
| | **1.** BI completo | 0,435 | 0,644 | 0,118 | 1 | 71 | 0 | 0,532 | 0,539 | 0,431 |
| **com BC** | **2.** EP=0,41 \| max=21 | 0,435 | 0,639 | -0,112 | 0,977 | 14 + 3 itens iniciais | 46 | 0,567 | 0,575 | 0,463 |
| | **3.** MR | 0,059 | 0,891 | -0,487 | 0,451 | 36 + 3 itens iniciais | 14 | 0,614 | 0,679 | 0,463 |
| | **3.** IE | 0,215 | 0,784 | -0,408 | 0,468 | 16 + 3 itens iniciais | 13 | 0,615 | 0,654 | 0,510 |
| **sem BC** | **4.** EP=0,41 \| max=21 | 0,436 | 0,629 | -0,026 | 0,981 | 15 + 3 itens iniciais | 47 | 0,562 | 0,569 | 0,458 |

Fonte: Elaborada pela autora.

Nesses casos, muitos itens do BI não foram utilizados no teste (47 e 46 itens, em torno de 65%), o que significa que os respondentes receberam muitos itens comuns. Esse fato é bastante discutido em Barrada et al. (2009), onde os autores mostram que indivíduos nos extremos recebem praticamente os mesmos itens quando nenhum método de controle da taxa de exposição condicional ao nível do traço latente é estabelecido, sendo este um fator muito preocupante para testes de alto impacto.

Na regra 3 (Tabela 7), onde a taxa máxima de exposição foi estabelecida em 0,4, nota-se que houve uma legeira redução na precisão das estimativas, aumento do RMSE e do viés. Entretanto, a taxa de sobreposição de itens foi bem menor do que nos outros estudos sem esta restrição (0,451 e 0,468). Com a inserção desta restrição, a maioria dos itens do BI foram utilizados. Por outro lado, a maioria desses itens não correspondem ao nível do traço latente desses respondentes.

A partir desses resultados, fica evidente a carência do BI disponível para mensurar com precisão o traço latente de respondentes que estão localizados no extremo superior da escala, região esta com poucos itens. Consequentemente, o teste é praticamente o mesmo para todos os respondentes quando nenhuma restrição da taxa máxima de exposição é imposta, podendo levar rapidamente ao pré-conhecimento de itens. Esses indivíduos acabam acertando todos ou quase todos os itens, não permitindo a diferenciação entre eles.

A restrição de BC obriga que itens de todos os Módulos sejam aplicados, mesmo que os itens restantes no BI estejam distantes do traço latente provisório do respondente. Isso acaba prejudicando a precisão das estimativas. Por exemplo, se não há itens difíceis em determinado Módulo, o algoritmo acaba selecionado um item mediano daquele Módulo para ser aplicado. Por isso, o método sem restrição de BC mostrou-se mais eficiente.

### 5.1.1 Conclusões gerais dos estudos

A simulação com dados reais (denominadas de *post-hoc*) é um passo muito importante antes da aplicação do CAT em situações reais, pois permite que os desenvolvedores de CAT avaliem características importantes do sistema, tais como a seleção de itens e regras de finalização do teste (BJORNER et al., 2007; THOMPSON; WEISS, 2011). As decisões tomadas a patir das simulações terão impacto nos

resultados futuros do CAT e na necessidade de manutenção do BI, por isso, merecem atenção especial.

Esses estudos simulados possibilitaram investigar diversas variáveis que podem ser manipuladas em um CAT, a fim de obter o melhor *design* de teste para o BI disponível, dentro das restrições deste programa de testes, investigando os impactos nas estimativas dos traços latentes, bem como no uso dos itens do BI. Assim, após diversos estudos e seus resultados, optou-se por implementar o seguinte algoritmo para o CAT, aplicado em um curso de capacitação para profissionais da área da saúde:

- Três itens fixos para iniciar o CAT (*start* 4 – Estudo 3);
- Método EAP para estimação do traço latente provisório e final (Estudo 1);
- Método MFI de seleção de itens com restrição de balanceamento de conteúdo (sete Módulos do curso de capacitação) (Estudo 4);
- Regra de parada: EP = 41 ou até 21 itens (Estudo 2).

Essas regras geraram os melhores resultados dentre os métodos investigados. Como o teste não era de alto impacto, a restrição da taxa de exposição dos itens não foi implementada no CAT operacional. No entanto, conforme esperado, a inserção da restrição da taxa máxima de exposição dos itens influencia na obtenção de piores valores para as seguintes estatísticas: correlação entre os traços latentes verdadeiros e estimados, RMSE, viés, EP máximo, proporção de respondentes que atingiram a regra de parada (EP=0,41) e aumento no comprimento médio do teste, em comparação aos resultados sem esta restrição.

Por outro lado, apresenta menor taxa de sobreposição de itens e uso mais homogêneo dos itens do BI, características importantes para a segurança dos testes de alto impacto. Consequentemente, as respostas são mais bem distribuídas entre os itens do BI, fato importante para a utilização dos procedimentos apresentados na FASE 2 da sistemática proposta, onde é feita a verificação de pré-conhecimento ou *drift* de itens que não são expostos com muita frequência.

Sabe-se que o BI utilizado é pequeno, o que causa grande impacto em todas essas estatísticas investigadas, principalmente pela falta de itens no extremo superior da escala e de diferentes níveis de dificuldade dos itens em todos os Módulos abordados no curso de capacitação, evidenciando a necessidade de calibração de novos itens para melhorar a precisão das estimativas dos traços latentes e obter CATs de melhor

qualidade. Além disso, salienta-se que, com este BI inical disponível para o CAT, seria inviável sua utilização em testes de alto impacto, uma vez que haveria a superexposição de muitos itens já no início de sua implementação.

Por fim, o Estudo 6 mostrou que respondentes com elevados níveis do traço latente receberam muitos itens em comum e isso pode torná-los conhecidos rapidamente. Desta forma, seria interessante utilizar um método de controle da taxa de exposição condicional aos níveis do traço latente. Para isso, mais itens precisam estar disponíveis nos níveis extremos.

## 5.2 APLICAÇÃO DO CAT E MANUTENÇÃO DO BI

Os resultados são apresentados em conjunto com a descrição da implementação de algumas etapas da sistemática para a manutenção do BI, que foi desenvolvido para avaliar profissionais de um curso de capacitação na área da saúde. Também, apresentam-se análises dos dados obtidos das aplicações dos CATs em duas edições de testes.

### 5.2.1 Primeira edição do CAT – dezembro de 2015

FASE 1 – CALIBRAÇÃO DE NOVOS ITENS

Ao iniciar a aplicação do CAT aos respondentes, novos itens estavam disponíveis para serem calibrados, iniciando a FASE 1 de manutenção do BI. Nesta edição, 827 profissionais participaram do teste.

**(A) Quantos itens de pré-teste aplicar e qual o tamanho da amostra?**

O CAT possui comprimento variável (EP=0,41), sendo o comprimento máximo possível de 21 itens. Com base nisso, optou-se por inserir quatro itens novos para serem pré-testados. Assim, o respondente poderia receber até 25 itens, não aumentando muito o comprimento do teste. Este número de itens de pré-teste corresponde a 19% do comprimento máximo do CAT operacional ou 16% de itens de pré-teste no teste, estando de acordo com Zheng (2014).

Para a calibração não-adaptativa dos itens, optou-se por apresentar os quatro itens novos a todos os respondentes, uma vez que era esperado entre 700 e 1.000 respondentes. Desta forma, se dois blocos distintos de itens de pré-teste fossem utilizados, em torno de 350 respostas a cada item

seriam possíveis, e esse tamanho de amostra poderia gerar imprecisões nas estimativas dos parâmetros dos itens.

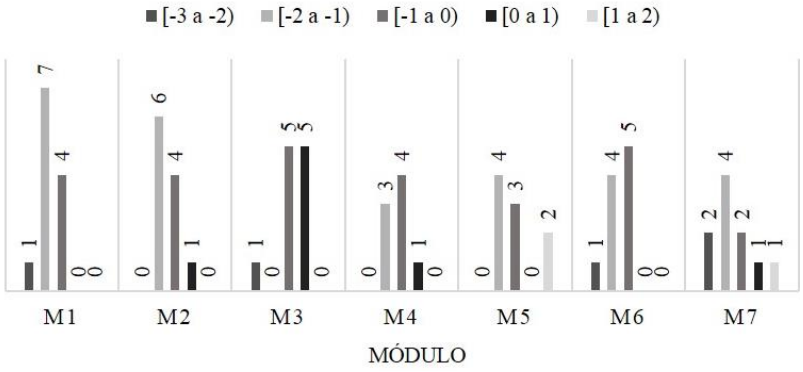
Como a exposição dos itens não era um problema neste teste e devido à urgência em aumentar o BI, a possibilidade de ter que aplicá-los novamente numa próxima edição para obter uma amostra maior de respondentes não seria interessante.

**(B) Quais itens de pré-teste devem ser calibrados?**

Vários itens já elaborados por especialistas estavam disponíveis para serem pré-testados. No entanto, era preciso dar prioridade para regiões da escala com pouca informação. Algumas análises descritivas foram efetuadas para auxiliar nesta decisão.

No BI para CAT, dentre os 71 itens disponíveis, somente 11 itens possuíam parâmetros de dificuldade acima da média ($b$=0) e nenhum item acima de $b$ = 2. Isto implica em sérios problemas na estimação dos traços latentes dos respondentes localizados no extremo superior da escala. Além disso, este problema se agrava quando a restrição de BC é imposta porque o algoritmo busca dentro de cada Módulo, itens nos níveis dos respondentes para serem apresentados. A Figura 21 mostra os níveis de dificuldade dos itens separados em cinco intervalos nos sete Módulos.

Figura 21 – Número de itens nos diferentes níveis de dificuldade por Módulo de conteúdo.




Fonte: Elaborada pela autora.

Observa-se, na Figura 21, que o Módulo 7 (M7) é o único que possui itens, mesmo que poucos, nos cinco níveis de dificuldade

avaliados, na escala N(0,1). Os níveis [-2 a -1) e [-1 a 0) possuem maior quantidade de itens, 28 e 27 itens, respectivamente. Os Módulos 1, 2, 3, 4 e 6 não possuem itens no nível [1 a 2).

Essa análise permite a visualização dos pontos da escala que precisam ser acrescentados mais itens. Quando os itens de pré-teste são selecionados para calibração junto ao CAT operacional, eles devem ser escolhidos com o objetivo de tentar suprir esses espaços vazios na escala. Assim, com base nos itens já elaborados e disponíveis para serem calibrados (itens por Módulos e níveis de dificuldade definido *a priori* pelo especialista), selecionaram-se quatro itens classificados como difíceis pelos especialistas, uma vez que se tem por objetivo calibrar itens nos níveis superiores da escala. Esses itens pertenciam ao Módulo 1 (M1).

**(C) Definição do *design* de calibração**

Para calibração dos itens, utilizou-se o *design* não-adaptativo. Uma vez definido pelos especialistas quais e quantos itens deveriam ser pré-testados, eles foram aplicados junto ao CAT operacional, para todos os respondentes (827), conforme regras apresentadas a seguir.

Esses itens não foram computados no traço latente do respondente e após atingirem a regra de parada para calibração (aplicar o item aos 827 respondentes), os dados foram armazenados para serem calibrados utilizando o *software* Bilog.

**1. Local de inserção dos itens de pré-teste no CAT**

Os itens foram aplicados conforme a seguinte regra: três itens fixos para todos os respondentes para iniciar o teste e a cada dois itens operacionais, um item de pré-teste era aplicado; isso continuou até que os quatro itens de pré-teste fossem administrados.

Após a aplicação dos itens de pré-teste, o respondente seguia seu teste operacional até atingir a regra de parada predefinida para o CAT. Assim, o indivíduo precisava responder, pelo menos, 11 itens operacionais para que todos os itens de pré-teste fossem aplicados pela regra dada. Caso o indivíduo terminasse o teste operacional antes de serem administrados todos os itens de pré-teste seguindo esta regra, os itens de pré-teste restantes eram apresentados no final do CAT operacional, antes de repotar o traço latente ao respondente.

Optou-se por esta regra de apresentação dos itens de pré-teste para calibração porque o teste possui comprimento variável e, conforme identificado por meio de simulação no Estudo 4 (Tabela 5), alguns

respondentes poderiam terminar o teste com nove itens operacionais e, em média, com aproxidamente 15 itens. Assim, itens de pré-teste foram "misturados" aos itens operacionais no CAT.

Além disso, como os itens de pré-teste foram apresentados a todos os respondentes (calibração não-adaptativa), não era necessário obter maior precisão das estimativas dos traços latentes para melhorar a seleção de itens de pré-teste, como ocorre na calibração adaptativa de itens, em que autores como Ali e Chang (2014), van der Linden e Ren (2015) e Zheng (2014) sugerem a aplicação desses itens ao final do teste operacional.

**2. Método de estimação dos parâmetros dos itens de pré-teste**

Os quatro itens de pré-teste foram aplicados a todos os respondentes para obter uma amostra razoável de respostas, durante o CAT operacional, e foram armazenadas para calibração utilizando o *software* BILOG (ZIMOWSKI et al., 2003), cujo método de estimação implementado é o MML e estimador EAP. Uma vez definido o método de estimação, ele deve ser mantido ao longo do tempo.

Para colocar os itens de pré-teste na mesma escala de medida dos itens do BI, esses itens novos foram calibrados junto com os itens operacionais do BI, os quais são considerados fixos e servem de ligação para estimar apenas os parâmetros dos itens de pré-teste.

**(D) Critérios para avaliar a qualidade dos itens pré-testados**

- **Precisão e bons parâmetros**

Nesta etapa, é preciso analisar o EP dos itens e o valor dos parâmetros estimados, visando identificar possíveis problemas nos itens. Se algum item mostrar-se suspeito, os especialistas de conteúdo podem auxiliar na identificação do problema, bem como na tomada de decisão de excluir o item.

Obteve-se 827 respondentes ao CAT operacional e, consequentemente, aos itens de pré-teste. Dos quatro itens submetidos à estimação, conforme apresentado na Tabela 8, um item apresentou problemas nos seus parâmetros e foi enviado para os especialistas darem seus pareceres sobre a qualidade do item.

Este item apresentou elevado EP da estimativa de discriminação do item (0,557), bem como elevado parâmetro para a probabilidade de acerto ao acaso (0,5). O *feedback* dos especialistas foi favorável a

exclusão do item por apresentar problemas em suas alternativas de resposta, ou seja, distratores não apropriados e confusão de conceitos.

O item I74 também apresentou elevado EP para $a$ = 2,44 (0,544). No entanto, a maginitude do parâmetro é elevada, o que pode estar influenciando na magnitude do EP, uma vez que os especialistas não identificaram problemas no item. Assim, os itens I72, I73 e I74 foram enviados para o BI operacional e a FIBI foi atualizada, totalizando 74 itens no BI.

Tabela 8 – FASE 1: Estimativa dos parâmetros dos itens de pré-teste aplicados na 1ª edição de testes e respectivos erros padrão.

| **Item** | ***a*** | ***b*** | ***c*** |
|---|---|---|---|
| I72 | 0,946 (0,167) | -1,331 (0,366) | 0,295 (0,106) |
| I73 | 2,011 (0,324) | 0,342 (0,073) | 0,079 (0,027) |
| I74 | 2,44 (0,544) | 0,298 (0,098) | 0,331 (0,035) |
| Excluído | 1,912 (0,557) | 0,239 (0,164) | 0,500 (0,045) |

Fonte: Elaborada pela autora.

Notou-se ao observar as estimativas dos parâmetros dos itens na Tabela 8, que a classificação dos especialistas quanto à difuldade dos itens são diferentes das obtidas na calibração, ou seja, não obteve-se nenhum item com elevado nível de dificuldade, somente $b < 0,35$. Desta forma, não foi possível aumentar a informação no extremo superior da escala de medida, conforme se pretendia.

Obviamente, neste caso, como o teste não é aplicado em um ambiente controlado, acaba interferindo na dificuldade em conseguir itens nos níveis elevados da escala. Mas, não deixa de ser impotante destacar a dificuldade de se elaborar itens para determinado nível de dificuldade e que atenda a determinado requesito de conteúdo.

Este fato dificulta a substituição de itens com a mesma funcionalidade no BI, os quais se tornam superexpostos, pré-conhecidos ou obsoletos ao longo do tempo. Enfatiza-se, assim, a importância de uma análise mais aprofundada dos itens que atingem a taxa máxima de exposição em CATs de alto impacto, antes da sua efetiva exclusão.

- **Pressupostos da TRI**

Comumente é assumido que, se os dados se ajustam adequadamente ao modelo unidimensional da TRI utilizado para calibração dos itens, há um fator dominante do traço latente responsável pelo itens do BI. Violar o pressuposto da unidimensionalidade pode levar

a desajustes na estimação dos parâmetros ou erros padrão elevados (DEMARS, 2010). Assim, a análise dos pressupostos da TRI e da existência de DIF são importantes.

A análise fatorial de informação completa para o BI inicial já desenvolvido mais os três itens novos, totalizando 74 itens, mostrou variância explicada pelo primeiro fator de 42,9%. O item I74 (Tabela 8), que apresentou apresentou EP considerável na estimação, também foi investigado quanto à sua dimensionalidade. No entanto, demostrou elevada carga fatorial no primeiro fator (0,84).

- **Detecção de DIF**

Esta etapa de verificação não foi efetuada neste trabalho, pois não foram repassadas informações sobre o perfil dos respondentes de modo a possibilitar as comparações entre grupos de respondentes.

FASE 2 – ETAPA 1: MONITORAMENTO DA EXPOSIÇÃO DOS ITENS

Conforme já mencionado, não foi inserido controle da taxa de exposição dos itens, pois o teste originalmente não é de alto impacto. Além disso, o BI real deste estudo é muito pequeno e não comportaria a inclusão de mais esta restrição.

Conforme a literatura (STOCKING, 1994; SEGALL, 2005; DAVEY, 2011; OZYURT et al., 2012; WAY, 1998), tamanhos ideais do BI para CAT, geralmente variam de 5 a 10 vezes o número de itens a serem aplicados para os respondentes. Portanto, abordando o comprimento máximo de 21 itens do CAT operacional, seria adequado ter entre 105 a 210 itens no BI, pelo menos. Também, não há itens suficientes nos diferentes níveis de dificuldade por Módulo, fato que provavelmente os tornariam rapidamente conhecidos.

No entanto, discussões acerca da exposição desses itens nas edições de testes serão apresentadas, bem como análises descritivas sobre o tempo de resposta, cuja fonte de informação pode ser utilizada paa monitoramento de mudanças ao longo do tempo e para buscar indícios de pré-conhecimento de itens.

**1. Análise da taxa de exposição geral dos itens do BI**

A Figura 22-I representa a taxa de exposição dos 74 itens (71 + 3 novos) obtidas desde a calibração dos itens, onde o item é exposto pela primeira vez, até a finalização da 1ª edição dos CATs. Observa-se dois

grupos bem distintos quanto à magnitude das taxas, sendo os primeiros itens a compor o BI, os mais expostos. Este fato destaca a importância de expor o menos possível os itens desde a fase de calibração. A calibração adaptativa on-line pode ser uma boa opção neste caso.

Do total de itens, 25 apresentaram uma taxa de exposição acima de 0,4 (taxa máxima definida no Estudo 5), que corresponde a 33,8% dos itens. Os três itens com maior taxa de exposição na Figura 22-I e Figura 22-II foram I16 (taxa = 0,933), I28 (0,913) e I07 (0,880) que pertencem aos Módulos M4, M6 e M2, respectivamente. Os itens I16 e I28 foram utilizados como itens iniciais fixos nesta edição do CAT, o que influenciou na obtenção de uma elevada taxa, inclusive, foram utilizados como itens fixos no desenvolvimento do BI, em junho de 2015, para o teste não-adaptativo, sendo aplicado a todos os respondentes.

A Figura 22-II mostra que, tanto itens com alta discriminação quanto baixa foram administrados. Entre os itens com alta discriminação, alguns ainda possuem baixa taxa de exposição. A Figura 22-III apresenta a taxa de exposição dos itens em relação aos parâmetros do dificuldade. Observa-se que os três itens com $b > 1$ (2 itens de M5 e 1 item de M7) tiveram alta taxa de exposição, variando entre 0,6 e 0,7. Itens com níveis de dificuldade no extremo inferior da escala apresentaram baixa taxa de exposição.

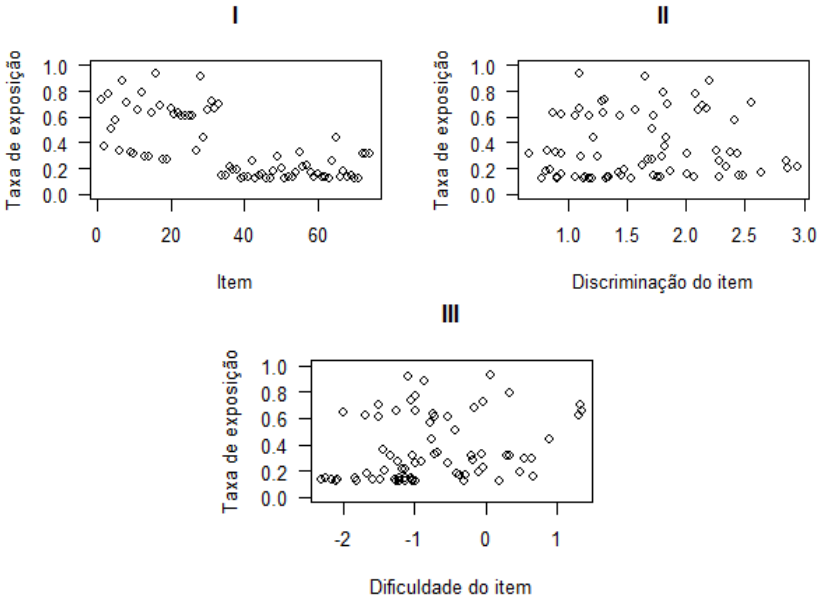
Isso indica que poucos respondentes que participaram dos testes estavam localizados neste nível (abaixo de -2); já os itens com maior taxa de exposição possuem níveis de dificuldade entre -1,5 e 0, região da escala com mais informação (Figura 17) e onde há mais respondentes (Figura 28 – IV).

Obviamente, se a taxa de exposição tivesse sido estabelecida no algoritmo computacional do CAT, essas taxas teriam sido diferentes para todos os itens do BI. Porém, de forma geral, 29 itens apresentaram taxa de exposição abaixo de 0,2 (39,2%), taxa esta, muitas vezes utilizada como taxa máxima em testes de alto impacto e, 49 itens apresentaram taxa de exposição abaixo de 0,4 (66,2%).

Com base nesta informação, se fosse considerado a exclusão dos itens que atingem a taxa máxima predefinida, sugestão dada por diversos autores, muitos itens teriam de ser eliminados, prejudicando ainda mais a precisão das estimativas dos traços latentes e impossibilitando o uso desses testes para avaliar os profissionais do curso de capacitação. Por isso, novamente, ressalta-se a importância da execução das demais etapas da sistemática proposta neste trabalho para tomar decisões mais assertivas

quanto à real ameaça de um item para avaliação e sua posterior eliminação do BI.

Figura 22 – Taxa de exposição dos 74 itens do BI (após 1ª edição do CAT).



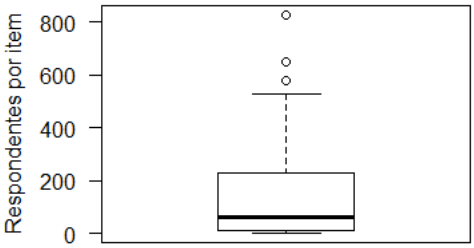

Fonte: Elaborada pela autora.

## 2. Análise da exposição dos itens no CAT operacional – 1ª edição

Considerando apenas os resultados da aplicação do CAT para verificar a utilização dos itens que estavam disponíveis, a Figura 23 mostra a dispersão entre o número de respostas obtidas para os itens do BI. Assim, tem-se que: dos 71 itens que estavam disponíveis no BI inicial para o CAT operacional, três itens iniciais foram apresentados aos 827 respondentes (*outlier*); nove itens não foram utilizados nenhuma vez nessa edição do CAT e, para os demais itens, a amostra de respondentes variou de 1 a 647 (*outlier* – I07), com média de 182,9 e mediana de 63 respondentes por item.

Caso a restrição da taxa de exposição tivesse sido imposta no algoritmo, muitos itens não teriam sido selecionados para admistração com tanta frequência como ocorreu para alguns deles.

Figura 23 – Variação do número de respondentes por item no CAT operacional: 1ª edição.




Fonte: Elaborada pela autora.

### 3. Análise do tempo de resposta ao item e de teste no CAT

Nesta seção deveriam ser efetuadas análises para tentar detectar o pré-conhecimento de itens que atingiram taxa de exposição superior a 0,4 (25 itens) ou dos outros itens que não atingiram esta taxa, mas que podem ter se tornado pré-conhecidos, dependendo da sua localização na escala (se o método de controle da exposição do item não é condicional ao traço latente) utilizando a informação de RTs ou um dos métodos apresentados na seção 3.1.2.1.2.

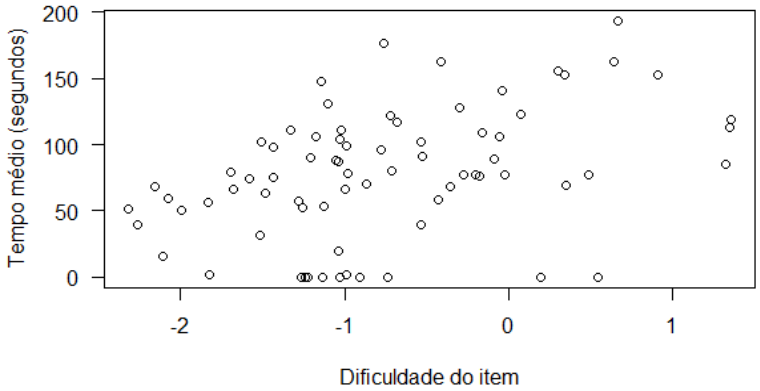
Devido às peculiaridades do CAT, não foi implementado nenhum procedimento para verificação de itens pré-conhecidos. Por outro lado, análises descritivas relacionadas ao tempo de resposta ao item e ao teste para o CAT foram efetuadas. Essas informações dos RTs são muito importantes para detecção de itens comprometidos em testes de alto impacto.

Para os itens com baixa taxa de exposição, pode-se monitorar os RTs para verificar se houve mudança significativa desde sua calibração. Para o item não aplicado ou muito pouco aplicado nos CATs, sugere-se o estabelecimento de um prazo razoável de uso para esses itens (definido pelo especialista de conteúdo) e caso mantenham-se com baixa taxa de exposição, pode-se aplicá-los em edições posteriores como sendo um item de pré-teste para verificar se houveram alterações nos parâmetros dos itens (DPI), conforme sugerido por Guo (2016), na Fase 2 (Etapa 2) da sistemática.

A Figura 24 representa o tempo médio gasto no item em relação ao parâmetro de dificuldade, na 1ª edição. Assim, ao observar o tempo para os 74 itens, nota-se uma tendência de que, quanto mais difícil o item é,

mais tempo os respondentes levam para finalizá-lo. Há uma correlação positiva, mas não tão forte entre essas variáveis (0,413).

Figura 24 – Média do tempo de resposta ao item em relação à dificuldade: 1ª edição.




Fonte: Elaborada pela autora.

No estudo de van der Linden, Scrams e Schnipke (1999), a dificuldade do item e o tempo de resposta ao item foram positivamente correlacionados. Outros estudos também destacam que itens mais difíceis geralmente requerem mais tempo dos respondentes (VAN DER LINDEN; SCRAMS; SCHNIPKE, 1999; CHANG; PLAKE; FERDOUS, 2005).

Observa-se, também, que nove itens aparecem com tempo médio igual a zero na Figura 24. Esses itens não foram utilizados no CAT. O tempo médio de 2 segundos foi identificado para dois itens ($b$ = -0,986 e -1,821), os quais foram respondidos somente por um indivíduo, sendo o mesmo nos dois itens. Este respondente provavelmente forneceu respostas aleatórias rápidas, dado que elas foram incorretas, o que não caracteriza um possível comportamento de pré-conhecimento dos itens, conforme destacado na literatura (WISE, 2014; VAN DER LINDEN; VAN KRIMPEN-STOOP, 2003). O maior tempo médio foi de 193,4 segundos (3,22 minutos), cujo item possui $b$ = 0,66 e foi respondido por 84 indivíduos, obtendo 58,3% de respostas corretas.

A Tabela 9 apresenta um resumo do CAT operacional quanto ao tempo de teste, comprimento do CAT e número de acertos. Os dados dos itens de pré-teste não estão incluídos nessas análises. Em média, os

respondentes levaram 24 minutos para finalizar o teste operacional, que teve como comprimento médio 13,33 itens e a média de acertos nos testes foi 8,57 itens.

Tabela 9 – Estatísticas gerais do CAT operacional (1ª edição).

| | **Mínimo** | **Mediana** | **Média** | **Máximo** | **Desvio padrão** |
|---|---|---|---|---|---|
| Tempo de teste (seg.) | 78 | 1285 | 1445 | 4533 | 774 |
| Comp. CAT | 8 | 12 | 13,33 | 21 | 3,64 |
| Nº acertos | 2 | 8 | 8,57 | 21 | 4,11 |

Fonte: Elaborada pela autora.

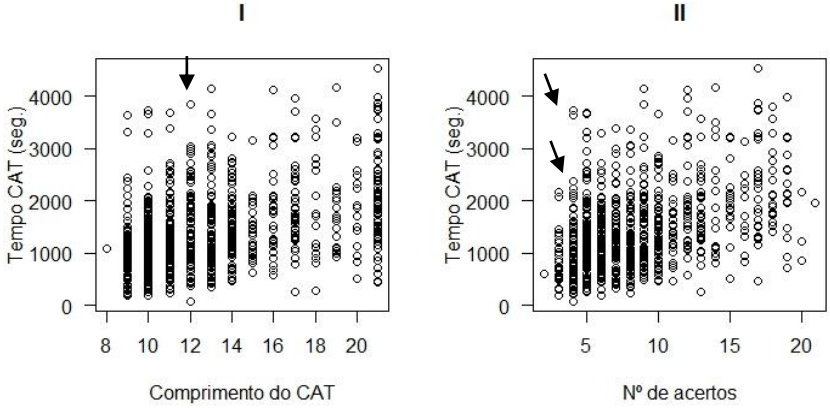
A Figura 25 apresenta a dispersão do tempo de teste (em segundos) condicionado aos diferentes comprimentos dos CATs, bem ao número de itens respondidos corretamente em cada CAT. Observa-se uma grande variabilidade nos tempos de teste para essas variáveis investigadas. Esta variação no tempo pode estar relacionada com diferentes níveis do traço latente (conforme Figura 28 – I).

Na Figura 25-I, por exemplo, 98 indivíduos que tiveram um comprimento de teste operacional igual a 12 itens, obtiveram seus tempos de teste variando entre 78 segundos e 3.834 segundos (aproximadamente 1h 04 min.), com média de 1426,85 segundos (23,8 minutos), traços latentes variando de -1,93 a 0,16 e número de itens respondidos corretamente variando entre 4 e 9 itens. O indivíduo que respondeu aos 12 itens em 78 segundos (média de 6,5 segundos por item), acertou 4 itens (33,3% do teste). Neste caso, possível pré-conhecimento de alguns itens poderia ser investigado.

Em testes de alto impacto, pode ocorrer um elevado tempo de teste, seguido de respostas incorretas, o que indicaria a tentativa de memorização de alguns itens para posterior divulgação. Na Figura 25-II, por exemplo, dois indivíduos que acertaram 4 itens, levaram 3.640 e 3.726 segundos (1h 02 min.) para encerrar o teste, cujo comprimento foi de 9 e 10 itens, respectivamente, e traços latentes de -0,918 e -1,415.

Esses dados provavelmente não indicam este comportamento, por mais que tenham sido bem distinto dos demais respondentes que acertaram 4 itens, uma vez que os dois respondentes acertaram 40% do teste ou mais. Neste caso, o tempo elevado pode indicar outro tipo de comportamento, dado que o teste não é de alto impacto, desde distração até a busca por fontes de resposta.

Figura 25 – Variação no tempo de teste em relação ao comprimento e número de acertos na 1ª edição.




Fonte: Elaborada pela autora.

Também, verificou-se que esses dois respondentes gastaram, pelo menos, 1.272 segundos a mais do que o tempo gasto pelo respondente que se encontra no topo do grupo com tempo de teste mais concentrado (de 2.368 segundos), cujo comprimento do CAT variou entre 9 e 15 itens. A média do tempo de teste para quem acertou 4 itens foi de 1.065,96 segundos (17,8 minutos).

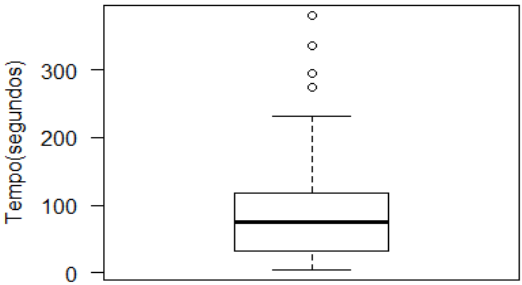
- **Tempo de resposta para itens não respondidos**

Dos 827 testes aplicados em dezembro de 2015, 32 respostas foram omitidas, não sendo um número expressivo. Esses itens não respondidos foram tratados como incorretos (conforme aviso dado ao respondente no momento do teste) para a seleção do próximo item no CAT operacional.

A não-resposta ocorreu em 20 diferentes itens do BI e a maior frequência de sua ocorrência no mesmo item foi três vezes, para 4 itens diferentes (cujo *b* varia entre -1,021 e 0,666). A Figura 26 representa a dispersão do tempo gasto nesses itens não-respondidos.

O tempo mínimo foi de 5 seg., tempo médio de 103,9 seg., mediana de 74,5 seg. e tempo máximo de 380 seg. Não houve um padrão para esses itens, porém, nota-se que foram tempos relativamente curtos quando comparados à média do tempo de resposta para outros itens, por exemplo, para os itens de pré-teste (média do tempo (geral) - Tabela 9).

Figura 26 – Boxplot do tempo gasto em itens não respondidos na 1ª edição.




Fonte: Elaborada pela autora.

- **Tempo de resposta para os itens de pré-teste (3 itens novos)**

Conforme apresentado nos procedimentos da Etapa 2 e destacado por Qian et al. (2016) e van der Linden e Guo (2008), o tempo que os indivíduos levam para responder aos itens na calibração pode ser usado para obter informações sobre possível pré-conhecimento dos itens em edições futuras, dado que o item está sendo exposto pela primeira vez. Portanto, algumas medidas descritivas sobre os três itens recentemente calibrados são apresentadas na Tabela 10.

Tabela 10 – Resumo do tempo de resposta para os 3 itens novos (em segundos).

| Item | % acerto | Tempo (geral) | | | Tempo (corretas) | | | Tempo (incorretas) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Mín. | Média | Máx. | Mín. | Média | Máx. | Mín. | Média | Máx. |
| I72 | 77,9% | 3 | 111,4 | 847 | 4 | 106,8 | 847 | 3 | 127,9 | 656 |
| I73 | 34,6% | 4 | 152,2 | 893 | 6 | 140,8 | 893 | 4 | 158,3 | 891 |
| I74 | 53,0% | 1 | 155,7 | 861 | 1 | 165 | 861 | 2 | 145,2 | 847 |

Fonte: Elaborada pela autora.

Observa-se, na Tabela 10, que em todos os casos houveram respostas rápidas (até 6 segundos). A porcentagem de respostas corretas para o item I72 foi bem acima dos outros itens. Para os itens I72 e I73 gastou-se mais tempo, em média, para respostas incorretas, estando de acordo com os resultados obtidos em estudos de Bergstrom, Gershon e Lunz (1994) e Hornke (2000). No entanto, para I74, mais tempo foi gasto, em média, para respostas corretas. O máximo de tempo gasto foi sempre maior para as respostas corretas.

Esses valores podem ser monitorados ao longo do tempo para identificar mudanças, em conjunto com os métodos mais sofisticados de diagnóstico de itens pré-conhecidos com base no RT, apresentados na seção 3.1.2.1.2.

A Figura 27 representa a distribuição do tempo de resposta desses itens (corretas - cinza e incorretas - branco) para monitoramento ao longo do tempo. Este gráfico pode fornecer evidências sobre o RT a ser considerado como resposta rápida em cada item, pela mudança "brusca" na frequência para respostas corretas para os tempos muito curtos.

Por exemplo, para I72 (Figura 27–I), a frequência de respostas corretas foi maior em todos os intervalos de tempo, inclusive para o intervalo de 0 a 39 segundos, havendo grande diferença na frequência desse intervalo de tempo, para a frequência de 40 a 59 segundos. Esta sobreposição das respostas corretas sobre as respostas incorretas é esperada, pois 77,9% dos respondentes acertaram o item.

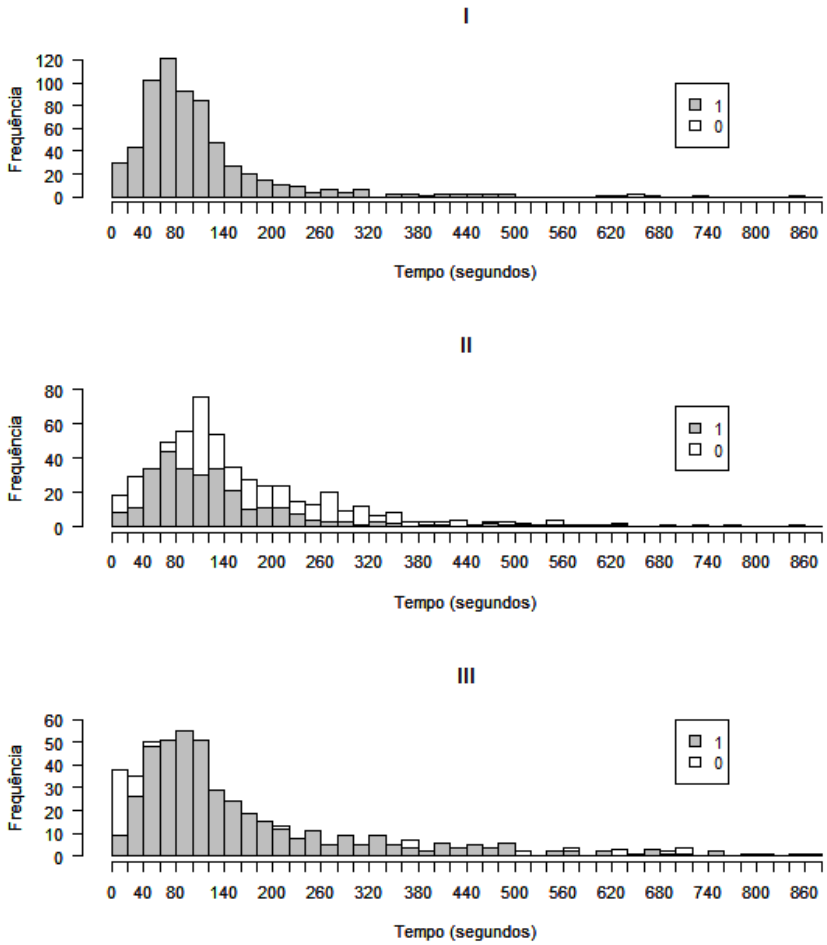
Para I73, nota-se que a frequência de respostas incorretas foi maior do que para respostas corretas na maioria dos intervalos de tempo. Além disso, no intervalo de 0 a 19 segundos, teve mais indivíduos respondendo rápido e errando o item, do que respondendo rápido e acertando. Essas informações serão comparadas com os resultados obtidos na 2ª edição do CAT para verificar se houve mudanças visíveis no gráfico de frequências.

## FASE 2 – ETAPA 2: VERIFICAÇÃO DE *DRIFT* DOS ITENS

Esta etapa não foi implementada neste trabalho. No entanto, como nenhuma restrição para a exposição dos itens foi imposta, alguns itens teriam respostas suficientes para serem recalibrados, enquanto outros, não. Como visto, teve itens que não foram utilizados no CAT, ou muito pouco utilizados, sendo impossível sua recalibração com os dados do CAT operacional.

Se isso se repetisse por várias edições, não atinjindo a taxa máxima de exposição predefinida, eles poderiam (mas não necessariamente) se tornar desatualizados e teriam que ser excluídos. Para recalibração nesses casos, eles teriam que ser considerados novamente como itens de pré-teste.

Figura 27 – Distribuição do tempo para respostas corretas e incorretas dos itens novos.




Fonte: Elaborada pela autora.

## RESULTADOS CONDICIONAIS À ESTIMATIVA DOS TRAÇOS LATENTES NO CAT OPERACIONAL – 1ª edição

A Figura 28 apresenta os principais resultados da aplicação dos CATs, condicionais aos traços latentes para o comprimento do CAT, número de acertos, EP das estimativas, tempo de teste, tempo médio por

item e a distribuição dos traços latentes desses respondentes. Na Figura 28-I, analisa-se o comprimento do teste para diferentes níveis do traço latente, cujos resultados mostram que, para as regiões em que há poucos itens no BI (extremos), os respondentes tiveram que responder aos 21 itens e a maioria respondeu até 14 itens.

O aumento no comprimento se torna evidente, assim como o EP das estimativas (Figura 28 - III), para $\hat{\theta} < -2$ e $\hat{\theta} > 0{,}5$. O maior EP obtido foi 0,575 e o menor EP igual a 0,351, estando de acordo com os resultados obtidos no estudo simulado (Estudo 4 – Tabela 5, método MFI). Além disso, 72 respondentes (8,7%) não atingiram a regra de parada EP = 0,41 e responderam a 21 itens, porcentagem menor do que os resultados obtidos na simulação (19%).

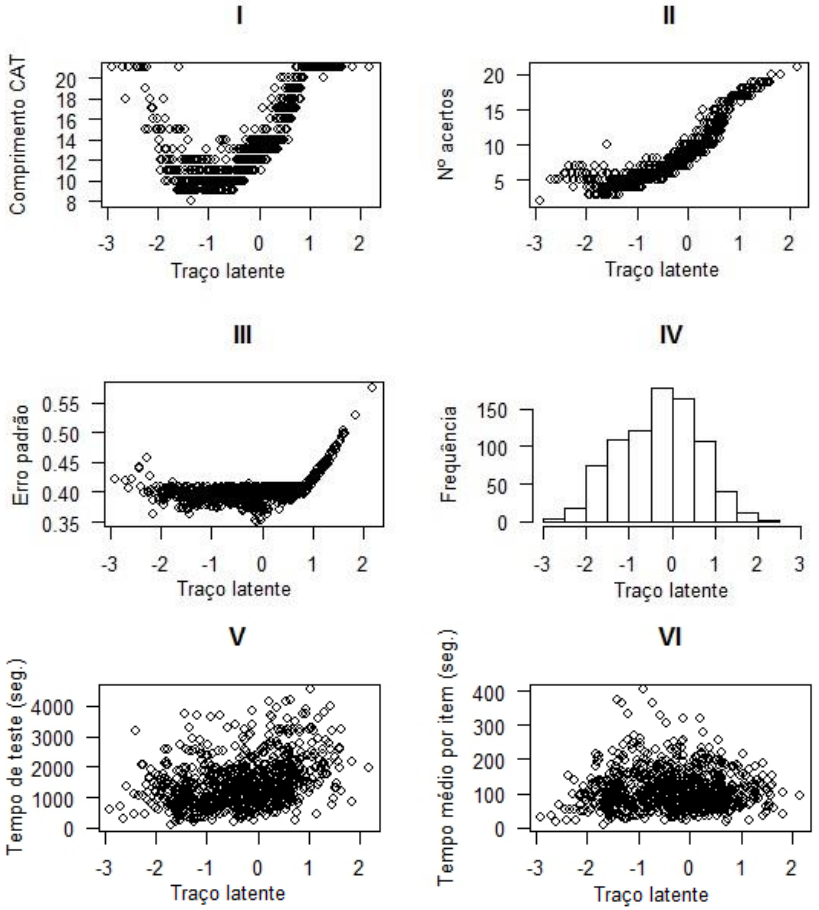
O histograma das frequências dos traços latentes pode ser visualizado na Figura 28-IV. As estimativas dos traços latentes variaram de -2,911 a 2,162, com média de -0,321, sendo que a maioria se encontra abaixo da média zero da escala de medida (60,9% dos respondentes).

Na Figura 28-II, observa-se o número de itens respondidos corretamente para os diferentes níveis do traço latente, apresentando uma correlação de 0,89. Assim, quanto maior o traço latente, maior o número de respostas corretas no teste. Nenhum respondente errou todos os itens; o mínimo de acertos foi igual a 2 para o pior desempenho. Em média, os respondentes acertaram 8,6 itens.

Segundo Nandakumar e Roussos (2004), em um CAT ideal espera-se obter o mesmo número de respostas corretas, independente do nível do traço latente do respondente. Este comportamento não ocorreu para este BI, obviamente, porque regiões da escala possuem níveis de informação diferenciada e diferentes quantidades de itens. No entanto, nota-se na Figura 28-II, que há uma semelhança no número de itens corretos para as regiões com mais informação no BI, como $-2 < \hat{\theta} < 0$.

Analisando a Figura 28-V e a Figura 28-VI, que tratam do tempo de teste e de resposta, nota-se que há uma certa tendência de que respondentes com maior traço latente e, consequentemente, que responderam a mais itens, apresentaram maior tempo de teste. Porém, os tempos de teste não foram altamente correlacionados com o traço latente (0,34), mas o tempo total de teste e a média dos tempos de resposta se correlacionaram bem (0,84). Resultados semelhantes também foram encontrados em Hornke (2000).

Figura 28 – Principais resultados condicionais às estimativas dos traços latentes da 1ª edição do CAT operacional.




Fonte: Elaborada pela autora.

Bergstrom, Gershon e Lunz (1994), Swanson et al. (1997) e Hornke (2000) não encontraram correlação entre as características do respondente e o tempo médio por item para prazos longos de realização do teste. Isso também ocorreu nestes CATs, onde os respondentes tinham tempo ilimitado para realizar o teste. Geralmente, indivíduos com baixo

traço latente não demoraram mais tempo para responder aos itens do que indivíduos com traço latente elevado.

### 5.2.2 Segunda edição do CAT – junho de 2016

Nesta edição, 878 profissionais participaram do teste. Como o algoritmo utilizado em edições futuras do CAT deve ser o mesmo das edições anteriores, na 2ª edição tem-se: 3 itens iniciais fixos, EAP para estimação do traço latente (provisório e final), MFI para seleção de itens e regra de parada igual a "EP=0,41 ou até 21 itens", com BC. Somente os três itens iniciais foram substituídos por outros três, com parâmetros dos itens semelhantes aos da 1ª edição.

Esta alteração é importante para testes de alto impacto, uma vez que os itens iniciais poderiam ter se tornado conhecidos, já que foram aplicados a todos os respondentes na edição anterior. As análises de pré-conhecimento e *drift* podem indicar se eles devem ser excluídos do BI, ou se podem ser mantidos por mais um tempo. Os principais resultados desta edição são apresentados a seguir.

FASE 1 – CALIBRAÇÃO DE NOVOS ITENS

Utilizou-se o *design* não-adaptativo de calibração on-line e, novamente, quatro itens de pré-teste foram aplicados seguindo as mesmas regras de inserção da 1ª edição. Esses itens foram considerados difíceis pelos especialistas e pertenciam aos Módulos M1, M2, M3 e M6. Esses Módulos não possuem itens nos níveis de dificuldade acima de 1 na escala (Figura 21).

A Tabela 11 apresenta os parâmetros desses itens e o EP de estimação, obtidos após a aplicação aos 878 respondentes. Observa-se que, igualmente à 1ª edição, não foi possível obter itens no extremo superior da escala. Os parâmetros *a* desses itens foram bons (acima de 0,7), assim como *b* e *c*; o EP foi adequado para o número de respondentes.

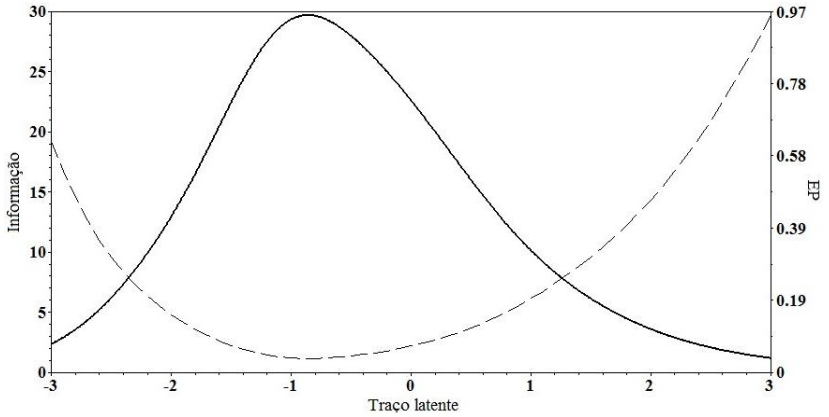
Posteriormente, a FIBI foi novamente atualizada, conforme mostra a Figura 29. Comparando a Figura 29, que teve um acréscimo de 7 itens, com a Figura 17 (BI inicial), praticamente não houve alteração na sua forma e precisão para mensurar os diferentes níveis do traço latente porque os itens novos estão em regiões da escala onde já havia mais informação.

Tabela 11 – Parâmetros dos itens pré-testados na 2ª edição e EP das estimativas.

| **Itens** | ***a*** | ***b*** | ***c*** |
|---|---|---|---|
| I75 | 2,303 (0,321) | 0,08 (0,064) | 0,089 (0,033) |
| I76 | 1,803 (0,238) | -1,693 (0,176) | 0,173 (0,078) |
| I77 | 1,954 (0,295) | -1,539 (0,163) | 0,166 (0,075) |
| I78 | 1,874 (0,259) | -1,306 (0,152) | 0,173 (0,076) |

Fonte: Elaborada pela autora.

Figura 29 – FIBI com 78 itens.




Fonte: Elaborada pela autora.

FASE 2 – ETAPA 1: MONITORAMENTO DA EXPOSIÇÃO DOS ITENS

**1. Análise da taxa de exposição geral dos itens do BI**

Nesta edição, I07 passou a ter a maior taxa de exposição geral (0,842), mantendo-se entre os três itens mais expostos. Posteriormente, tem-se o I01 (0,801), o qual foi utilizado como um dos itens inicias fixos do CAT nesta edição. Esses itens apresentam elevadas taxas e devem ser investigados para pré-conhecimento e *drift*, antes de sua exclusão por exposição excessiva.

Dos 78 itens do BI, 24 deles (30,8%) apresentaram taxa de exposição geral maior do que 0,4 e, consequentemente, 69,2% menor do que 0,4; 27 itens apresentaram taxa inferior a 0,2, que corresponde a 34,6%. Comparando as taxas gerais das duas edições, tem-se que,

conforme novos itens vão sendo acrescentados no BI, a quantidade de itens com taxa de exposição inferior a 0,2 ou inferior a 0,4, aumenta. Isso é esperado com a expansão do BI e é o desejo das organizações que aplicam testes de alto impacto: reduzir as taxas de exposição e, consequentemnte, de sobreposição de itens para obter níveis maiores de segurança do BI.

**2. Análise da exposição dos itens no CAT operacional - 2ª edição**

A Figura 30 mostra a distribuição do número de respondentes por item de uma forma geral (*boxplot*) e o número de respondentes condicionado aos níveis de dificuldade dos itens. Analisando o *boxplot* do número de respondentes para os 74 itens no BI, sete *outliers* apareceram: 878 (que corresponde aos três itens iniciais aplicados a todos) e outros itens que tiveram um número elevado de respondentes, isto é, 483, 546, 580, 587, 594, 639 (I07).

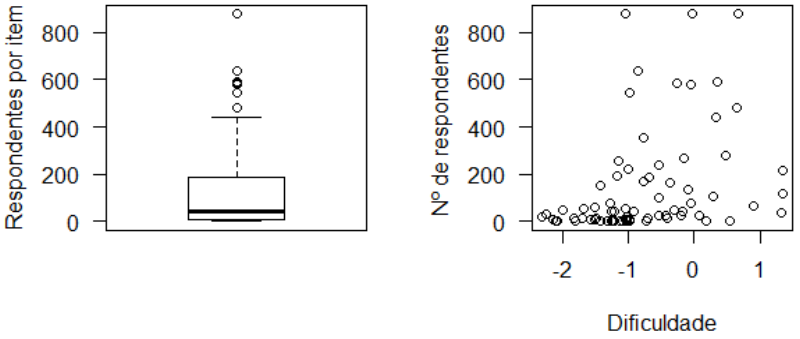
Esses itens pertencem a três Módulos diferentes e apresentam elevados parâmetros *a* (acima de 1,8) e diferentes valores de *b* (-0,979 a 0,64). Do total de itens,14 não foram apresentados a nenhum respondente no CAT operacional, cujos parâmetros *a* variaram de 0,77 a 2,07 e *b* variou de -2,1 a 0,54.

Identificou-se que o item I07 apareceu como *outlier* nas duas edições do CAT. Isso mostra que ele é um item que está sendo muito aplicado (pois não há restrição da taxa de exposição). Seus parâmetros são: $a = 2{,}1937$, $b = -0{,}8646$ e $c = 0{,}1311$ e percentence a M2, o qual há 4 itens disponíveis o intervalo de dificuldade [-1 a 0).

A média de respondentes por item foi de 144,3 e mediana de 42 para a 2ª edição. Embora mais *outliers* tenham sido identificados nesta edição, com mais itens no BI, a média e a mediana do número de respondentes por item foram menores.

Ao analisar o gráfico do número de respondentes de acordo com a dificuldade de cada item, na Figura 30, percebe-se que muitos itens com diferentes níveis de dificuldade são pouco aplicados. Provavelmente isso ocorre porque o algoritmo acaba selecionado sempre os melhores itens dentro de cada Módulo e, assim, quase sempre os mesmos acabam sendo selecionados, uma vez que não há controle da taxa de exposição dos itens. Além disso, o maior número de respondentes ocorre para os itens cujos níveis de dificuldade coincidem com os níveis em que há maior frequência de respondentes na escala de traços latentes (Figura 33 – IV), conforme esperado.

Figura 30 – Dispersão do número de respondentes para os itens no CAT operacional (2ª edição).




Fonte: Elaborada pela autora.

## 3. Análise do tempo de resposta e de teste no CAT

Os resultados da média do tempo de resposta ao item em relação à dificuldade do item foi muito semelhante aos resultados da 1ª edição (Figura 24). A Tabela 12 apresenta o resumo dos dados para o tempo de teste, comprimento do CAT e número de respostas corretas.

O número de acertos variou de 3 itens a 21 itens, sendo que apenas um indivíduo acertou todos os itens. O tempo máximo de teste (4.719 segundos ou, aproximadamente, 1h 18 min) foi obtido para 12 itens e 8 acertos. Também, houve baixa correlação entre comprimento do CAT e tempo de teste (0,3). Os valores médios para essas variáveis investigadas foram menores do que na 1ª edição, ou seja, em média, o tempo de teste, o comprimento e o número de acertos foram menores.

Tabela 12 – Estatísticas gerais do CAT operacional (2ª edição).

| | **Mínimo** | **Mediana** | **Média** | **Máximo** | **Desvio padrão** |
|---|---|---|---|---|---|
| Tempo de teste (seg.) | 43 | 1174 | 1313 | 4719 | 713,38 |
| Comp. CAT | 9 | 11 | 12,16 | 21 | 2,91 |
| Nº acertos | 3 | 7 | 7,85 | 21 | 3,26 |

Fonte: Elaborada pela autora.

Dois indivíduos apresentaram um tempo de teste muito rápido em comparação aos demais respondentes, obtendo 4 acertos. Um deles demorou 43 seg. para responder a 12 itens (média de 3,6 seg. por item) e

obteve 33,3% de êxito nas respostas. Esse comportamento poderia indicar pré-conhecimento dos itens ou respostas aleatórias.

Se o comportamento de respostas muito rápidas e acertos ocorressem para o mesmo item, por muitos respondentes, poderia indicar pré-conhecimento. O indivíduo que levou 49 seg. para responder a 21 itens (média de 2,3 seg. por item) e obteve 19% de êxito (abaixo de 0,2 - probabilidade de acerto ao acaso para cinco alternativas de respostas), provavelmente forneceu respostas aleatórias.

- **Investigando mudanças na distribuição dos RTs para corretas e incorretas**

As estatísticas dos RTs obtidas na fase de calibração para os itens I73 e I74, na 1ª edição do CAT (Tabela 10), foram comparadas com os RTs obtidos para esses itens quando aplicados adaptativamente aos respondentes nesta 2ª edição do CAT. O item I72 não foi administrado nenhuma vez nesta edição, portanto, não foi comparado; I73 foi aplicado a 443 respondentes e I74 foi aplicado a 106 respondentes.

Comparando os resultados da Tabela 10 com os resultados da Tabela 13, tem-se que a média de tempo para as respostas incorretas foi semelhante nas duas edições de testes para esses itens. Já a média de tempo para as respostas corretas foi um pouco diferente entre as duas edições, havendo uma redução no tempo médio de resposta na 2ª edição quando comparado à 1ª edição, para os dois itens.

Também, a porcentagem de respostas corretas aumentou na 2ª edição. Isso era esperado, uma vez que os itens são aplicados adaptativamente aos respondentes. Assim, I73 passou de 34,6% de respostas corretas para 52,1% e I74 passou de 53% para 73,6%. No entanto, ainda houve uma variabilidade nas respostas, mesmo sendo aplicado, teoricamente, no nível do respondente. Essa questão é bastante discutida na literatura quando se trata do uso dos dados do CAT operacional para a recalibração de itens e verificação de *drift*.

Tabela 13 – Resumo do RT no CAT operacional da 2ª edição para os dois itens que foram pré-testados na 1ª edição.

| Item | % acerto | Tempo (geral) | | | Tempo (corretas) | | | Tempo (incorretas) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Mín. | Média | Máx. | Mín. | Média | Máx. | Mín. | Média | Máx. |
| I73 | 52,1% | 6 | 143,6 | 849 | 6 | 129,1 | 683 | 9 | 159,5 | 849 |
| I74 | 73,6% | 12 | 141 | 553 | 20 | 140 | 553 | 12 | 143,7 | 530 |

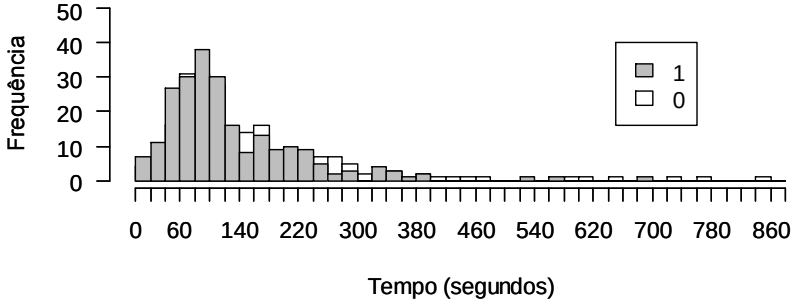
Fonte: Elaborada pela autora.

A Figura 31 apresenta a distribuição dos RTs para respostas corretas e incorretas para I73, aplicado no CAT operacional da 2ª edição de testes. Observa-se que a frequência de RT para as respostas corretas se sobrepôs ao RT incorreta em praticamente todos os intervalos de tempo, diferentemente do que ocorreu na 1ª edição do CAT para este item (Figura 27 – II).

Além disso, no intervalo de 0 a 19 segundos da Figura 31, teve mais indivíduos respondendo rápido e acertando, do que respondendo rápido e errando. O pico da distribuição de tempo para as respostas corretas aconteceu de 80 a 99 segundos, já na 1ª edição, ocorreu no intervalo de 60 a 79 segundos.

Esses resultados descritivos podem ser usados para monitoramento do item ao longo do tempo, que junto com os métodos de detecção de itens pré-conhecidos apresentados na Fase 2 (Etapa 1) da sistemática, podem fornecer informações mais conclusivas e fundamentadas sobre os itens e o que pode ser classificado como respostas rápidas para determinado item, com base no modelo para RTs proposto por van der Linden (2006).

Figura 31 – Distribuição dos RTs para o item I73 aplicado no CAT operacional.




Fonte: Elaborada pela autora.

## FASE 2 – ETAPA 2: VERIFICAÇÃO DE *DRIFT* DOS ITENS

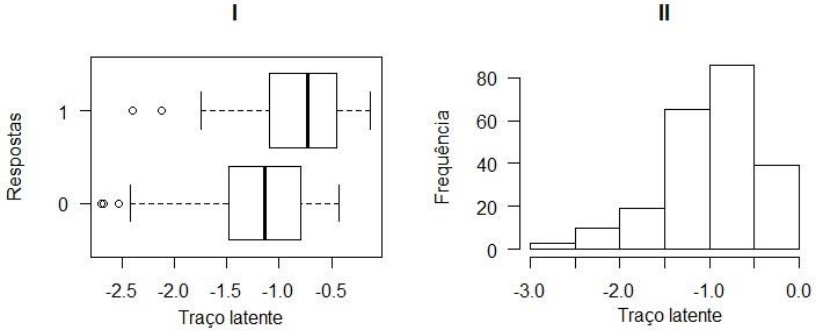
Esta etapa não foi implementada neste trabalho. No entanto, para um determinado item com certa frequência de administração no CAT operacional desta edição, investigou-se a distribuição dos traços latentes finais condicionais às respostas corretas e incorretas para o item (Figura

32-I), bem como a distribuição dos traços latentes finais para quem o item foi aplicado durante o processo (Figura 32-II).

Essa informação é importante em vista de uma possível tentativa de recalibração do item como método de verificação da ocorrência de *drift* dos parâmetros desse item. Este item (I11) foi apresentado a 222 respondentes e obteve 123 respostas corretas (55,4%); possui parâmetros $a$ = 1,576 e $b$ = -0,998.

Por meio do histograma na Figura 32-II, observa-se que o item foi aplicado com maior frequência para respondentes com traços latentes finais próximo ao nível de dificuldade do item. Por outro lado, nota-se uma grande variabilidade nas estimativas finais dos traços latentes para quem o item foi apresentado, mas todas elas estavam abaixo da média zero da escala. Isso provavelmente tem ligação com a ordem em que o item foi aplicado no teste, ou seja, as estimativas dos traços latentes tinham diferentes precisões.

Figura 32 – Distribuição do traço latente final dos indivíduos que responderam determinado item durante o CAT operacional.




Fonte: Elaborada pela autora.

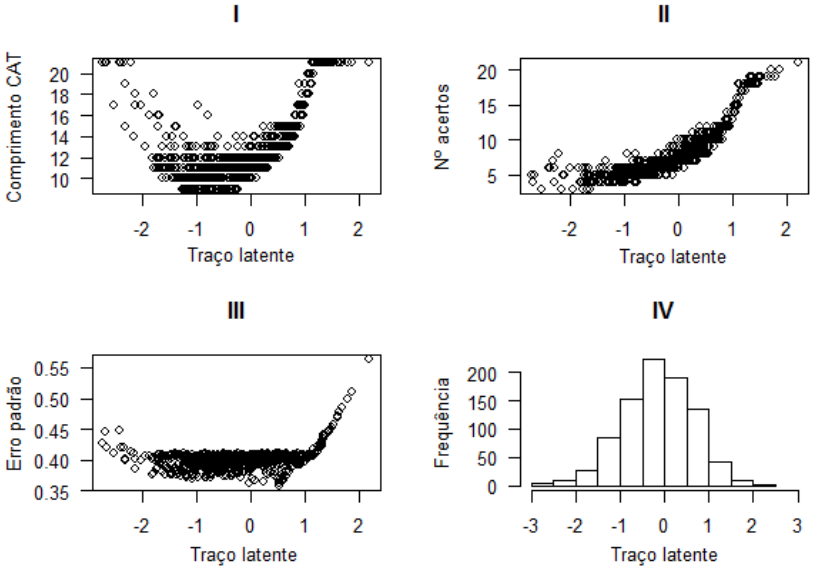
A dispersão do traço latente dos respondetes que erraram o item foi maior do que a dispersão dos traços latentes de quem acertou o item. Conforme esperado, o gráfico mostra que, de forma geral, quem acertou o item, possuía maior traço latente final do que os respondentes que erraram. No entanto, dois *outliers* podem ser observados para as respostas corretas, os quais possuíam níveis de traço latente bem abaixo do nível de dificuldade do item, supondo provável acerto ao acaso ou pré-conhecimento do item. Mas, para afirmar isso, mais estudos devem ser feitos.

Para recalibrar o item com precisão, Lu e Hambleton (2004) ressaltam que é preciso uma amostra razoavelmente grande e heterogênea. Esses resultados mostram que há uma variabilidade nos traços latentes e na frequência de erros e acertos. Porém, o item não foi aplicado a nenhum respondente com elevado nível de traço latente. Os impactos disso, podem ser investigados pela efetiva tentativa de recalibração no futuro.

## RESULTADOS CONDICIONAIS À ESTIMATIVA DOS TRAÇOS LATENTES NO CAT OPERACIONAL – 2ª EDIÇÃO

A Figura 33 apresenta uma síntese dos resultados, os quais mostram-se muito semelhantes à 1ª edição (em I, II e III), com uma pequena diferença no nível extremo inferior da escala de medida ( $\hat{\theta} < -2$). Na Figura 33-IV, que representa a distribuição de frequência dos traços latentes, nota-se uma mudança nesta distribuição quando comparada à 1ª edição.

Figura 33 – Principais resultados condicionais às estimativas dos traços latentes, na 2ª edição do CAT operacional.




Fonte: Elaborada pela autora.

A média dos traços latentes foi de -0,1475 e desvio padrão de 0,78, com -2,728 < $\hat{\theta}$ < 2,186. Logo, 57,1% dos respondentes obtiveram traço latente abaixo da média zero da escala de referência. Isso mostra que os respondentes desta edição apresentaram um desempenho melhor do que os respondentes da 1ª edição do CAT.

Na Figura 33-III, tem-se que 42 respondentes não atingiram a regra de parada EP = 0,41, que corresponde a 4,8%. O EP das estimativas variou de 0,358 a 0,563. Resultados que também estão dentro dos esperados pela simulação (Estudo 4 para MFI). Também, nesta edição, não houve correlação entre os traços latentes e a média de tempo por item e obteve-se uma baixa correlação entre o traço latente e o tempo de teste (0,21).

### 5.2.3 Conclusões gerais das aplicações dos CATs

A partir da Fase 1 da sistemática, não foi possível obter itens nas regiões da escala com pouca informação (extremo superior). Assim, ressalta-se a importância de calibrar mais itens, podendo até aumentar o número de itens de pré-teste por edição, visto que o comprimento médio foi em torno de 14 itens e há urgência em calibrar novos itens.

A taxa de exposição dos itens foi investigada, mostrando a importância da execução da Fase 2 para a manutenção do BI e tomada de decisão sobre a exclusão de itens que se tornam superexpostos e da inserção desta restrição para melhor uso dos itens e, consequentemente, melhorar a segurança do BI.

Sem esta restrição, alguns itens nunca foram apresentados, enquanto outros apresentaram elevadas taxas. No entanto, o BI real utilizado neste trabalho não suportaria a inserção desta restrição e não seria viável a utilização de um BI com essas características em testes de alto impacto.

Notou-se que, quando itens iniciais são fixos para todos os respondentes, eles apresentam elevadas taxas de exposição e podem se tornar facilmente conhecidos, prejudicando a avaliação. Porém, como ocorre para os demais itens do BI, um item pode atingir a taxa máxima de exposição em certo momento e, posteriormente, conforme outros itens vão sendo aplicados, esta taxa vai sendo reduzida. Logo, é importante verificar se realmente esses itens estão comprometidos antes de excluí-los.

Diferentes comportamentos dos respondentes foram investigados com base na resposta dada e no tempo de resposta, cujo foco foi

identificar discrepâncias para auxiliar na detecção de itens pré-conhecidos. Ressalta-se que o teste não é de alto impacto, logo, são apenas suposições com base em análises descritivas dos dados.

Essas análises auxiliam no monitoramento das informações ao longo do tempo. Métodos mais sofisticados para detecção do pré-conhecimento e *drift* não foram implementados. Encontrou-se grande variabilidade no tempo de teste quando analisado em relação ao comprimento do CAT e ao número de acertos, que pode estar associado aos diferentes níveis do traço latente ou à outros motivos, visto que o teste não foi aplicado em um ambiente controlado.

De forma geral, os CATs mostraram bom desempenho com um teste mais curto do que o teste não-adaptativo, principalmente para os respondentes localizados nas regiões com mais informação na escala de medida. O desempenho dos respondentes foi melhor na 2ª edição de testes. Além disso, os resultados das aplicações dos CATs mostraram-se em consonância com estudos da literatura quanto ao tempo de resposta ao item e de teste, assim como os resultados dos estudos simulados mostraram-se semelhantes com os obtidos na prática.

# 6. CONSIDERAÇÕES FINAIS

## 6.1 CONCLUSÕES

Este trabalho tentou responder a seguinte questão de pesquisa: **como manter um banco de itens para testes adaptativos computadorizados aplicados em avaliações de alto impacto ao longo do tempo?** Notou-se que o referencial sobre o tema apresentava limitações acerca dos procedimentos que devem ser adotados, quando eles devem ser executados e possíveis caminhos para auxiliar na tomada de decisões ao longo do tempo, ou seja, como executar esta tarefa na prática.

O estudo mostrou-se desafiador devido ao grande número de variáveis que podem ser manipuladas em um CAT, as quais impactam de forma diferenciada nos resultados de um teste. Por isso, uma literatura muito abrangente teve que ser consultada para que o principal objetivo deste trabalho fosse alcançado: desenvolver uma sistemática para auxiliar na manutenção do BI para testes adaptativos computadorizados aplicados em avaliações de alto impacto. A sistemática trata de questões básicas para a manutenção, sistematizando e sequenciando as ações necessárias para tal.

Comumente, restrições de exposição dos itens são impostas em CATs e, após atingirem uma taxa máxima predefinida, os itens são excluídos. No entanto, esta ação pode gerar vários problemas para as organizações de testes na prática. É importante ressaltar que itens frequentemente usados nem sempre estão comprometidos, assim como nada garante que os itens que não atingiram a taxa máxima de exposição não estejam comprometidos. Além disso, um BI relativamente grande ajuda a obter mais segurança, mas não é condição suficiente.

Problemas que surgem ao longo do tempo podem ser diagnosticados por meio da manutenção do BI. É natural que algumas alterações ocorram, porém, é preciso identificar e tratar os itens de forma adequada. Para isso, duas fases foram definidas para a manutenção.

A **Fase 1 -** CALIBRAÇÃO DE NOVOS ITENS envolve a definição de quantos e quais itens serão calibrados; definição do *design* de calibração, ou seja, do método de seleção de itens de pré-teste, do local de inserção do item de pré-teste no CAT, do método de estimação dos parâmetros do item e da regra de parada para itens de pré-teste; além da análise dos dados para a inclusão efetiva desses itens no BI.

A **Fase 2** engloba duas etapas de monitoramento dos itens. A **Etapa 1** - MONITORAMENTO DA EXPOSIÇÃO DOS ITENS tem por objetivo analisar as taxas de exposição e sobreposição de itens e identificar itens pré-conhecidos, sendo o tempo de resposta ao item uma importante fonte de informação nesta etapa. A **Etapa 2 -** VERIFICAÇÃO DE DRIFT DOS PARÂMETROS DOS ITENS envolve técnicas que visam identificar se itens tiveram seus parâmetros alterados ao longo do tempo e os motivos que levaram a esta ocorrência.

Esta sistemática visa a servir como um referencial para profissionais responsáveis pelo desenvolvimento e manutenção de CATs tomarem decisões sobre como proceder em determinadas situações, dando-lhes sugestões e apresentando métodos que podem ser aplicados em cada etapa de análise.

A sistemática desenvolvida pode ser aplicada, sobretudo, em avaliações de alto impacto, como em avaliações educacionais, testes de certificação ou licenciamento, seleção e classificação. Todavia, nada impede sua utilização, ou parte dela, em outros contextos. Por exemplo, em avaliações de baixo impacto, não faz sentido a implementação de procedimentos para detecção de itens pré-conhecidos. Portanto, esta etapa deve ser ignorada.

A manutenção é uma tarefa contínua, que deve ser executada em toda edição de testes. Na Fase 1, caso não existam itens elaborados para serem pré-testados, deve-se dar prioridade ao seu desenvolvimento. A Fase 2 é importante tanto para os itens que atingem a taxa máxima de exposição predefinida, quanto para itens que não atingem esta taxa ao longo do tempo. Os procedimentos adotados nas duas etapas da Fase 2 visam uma tomada de decisão mais assertiva de um possível comprometimento do item, baseando-se em um conjunto de evidências.

O desenvolvimento do CAT e a manutenção do BI deve ser realizada por uma equipe multidisciplinar de profissionais, tais como como psicometristas, especialistas de conteúdo e profissionais da área da computação. Para auxiliar na tarefa de iniciação de um teste adaptativo, o trabalho também forneceu diretrizes sobre as regras e *softwares* disponíveis para o desenvolvimento e implantação do CAT.

Diferentes estudos simulados mostraram as decisões tomadas para definição do *design* do CAT operacional para avaliação de profissionais de um curso de capacitação na área da saúde, que impacta diretamente na necessidade de manutenção (uso dos itens do BI) e na precisão das estimativas do traço latente após sucessivas edições de testes.

Com base nos resultados desses estudos, as seguintes regras foram implementadas para o CAT operacional: três itens iniciais com diferentes níveis de dificuldade (fácil, mediano e difícil), método EAP padrão (com distribuição *a priori* fixa N(0,1)) para estimar o traço latente provisório e final; método de máxima informação de Fisher para seleção de itens, com restrição de balanceamento de conteúdo; e comprimento variável do CAT, ou seja, o teste encerra após a obtenção de uma precisão de estimação igual a 0,41 ou, no máximo, 21 itens administrados.

Posteriormente, parte da sistemática de manutenção foi implementada junto com a aplicação do CAT operacional em duas edições de testes. Devido às restrições específicas do teste real utilizado neste trabalho e por não ser considerado de alto impacto, o controle da taxa de exposição não foi implementado no algoritmo. Porém, considerações acerca dos impactos que isso causaria no BI e análises dessas taxas após a aplicação do CAT operacional foram apresentadas.

Os resultados mostraram que o BI não é adequado para ser utilizado em avaliações de alto impacto, pois é muito pequeno (71 itens no BI inicial) e há pouca informação para mensurar os indivíduos localizados no extremo superior da escala, os quais receberiam praticamente os mesmos itens, favorecendo o pré-conhecimento de itens e prejudicando a avaliação de alto impacto. Devido ao tamanho do BI, se uma taxa máxima de exposição fosse imposta, teria grande impacto nas estimativas do traço latente, com redução considerável na precisão.

No CAT operacional, alguns itens mostraram elevada taxa de exposição, enquanto outros (nove - 1ª edição e 14 itens - 2ª edição) não foram administrados no CAT operacional. Obviamente, se o controle da exposição tivesse sido imposto no algoritmo, esses resultados teriam sido diferentes. Além disso, esses dados evidenciam a dificuldade em utilizar os dados do CAT operacional para recalibração de itens e detecção de *drift* dos parâmetros dos itens. Por isso, estratégias são fornecidas para lidar com esta condição de itens que não são expostos com muita frequência, mas que precisam ser investigados ao longo do tempo para identificar se os mesmos estão comprometidos.

Dados de tempos de resposta e de teste foram analisados e possíveis comportamentos de pré-conhecimento foram apresentados, bem como a relação dessas variáveis com outras, como comprimento do CAT, número de acertos e níveis do traço latente. Esses procedimentos auxiliam no monitoramento das informações para detecção de possíveis alterações nas características dos itens ao longo do tempo. No entanto, métodos mais

sofisticados devem ser implementados para auxiliar na tomada de decisão, conforme apresentado na sistemática.

Com o objetivo de suprir deficiências da escala, oito novos itens foram pré-testados e sete foram adicionados à escala de medida. Os resultados evidenciam a dificuldade de conseguir calibrar itens para atender pontos específicos de conteúdo e níveis de dificuldade. Por fim, os principais resultados quanto às estimativas do traço latente nas duas edições dos CATs são apresentados, os quais mostraram-se de acordo ou até melhor do que havia sido previsto nas simulações para definição do *design* do CAT.

Para finalizar, conclui-se que a sistemática fornece importantes orientações para o campo da prática de implementação e manutenção de CATs aplicados em avaliações de alto impacto, em que os desenvolvedores de testes não devem medir esforços para que os itens que compõem o BI sejam atualizados e para que itens comprometidos sejam, preferencialmente, excluídos para não comprometer a validade dos resultados da avaliação. Além disso, espera-se que novos itens sejam constantemente adicionados à escala para melhorar a precisão das estimativas do traço latente, levando à correção dos desvios identificados.

## 6.2 TRABALHOS FUTUROS

Para trabalhos futuros, sugere-se a implementação de métodos para detectar *drift* dos parâmetros dos itens para o BI utilizado no trabalho. Também sugere-se adaptar o método desenvolvido por Wise e Ma (2012) que trata da identificação de limites de tempo de resposta para comportamento solução e comportamento de acerto ao acaso para cada item, em CATs de baixo impacto, para o contexto de identificação de limites de tempo de resposta para itens comprometidos devido ao pré-conhecimento em CAT de alto impacto, cujo comportamento é respostas rápidas e corretas.

Outra sugestão de pesquisa futura é considerar a ordem em que os itens são apresentados no CAT para investigar se houve diferença no tempo de resposta para um item, por exemplo, quando são aplicados no início do teste e quando são apresentados adaptativamente próximos ao final do teste, bem como sua relação com o aumento da taxa de exposição do item. Além disso, sugere-se analisar o impacto de incluir respondentes que fornecem respostas rápidas para o item na calibração, ou então, excluí-lo do processo.

## REFERÊNCIAS

ABAD, F. J.; OLEA, J.; AGUADO, D.; PONSODA, VICENTE; BARRADA, J. R. Deterioro de parámetros de los ítems em tests adaptativos informatizados: estudio con eCAT. *Psicothema***,** v. 22, n. 2, p. 340-347, 2010.

ABEPRO. *Associação Brasileira de Engenharia de Produção*. 2015. Disponível em: http://www.abepro.org.br/indexsub.asp?ss=40. Acesso em: set. 2015.

AERA, APA, NCME. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington, DC: AERA, 1999.

ALI; U.S.; CHANG, H. H. *An Item-Driven Adaptive Design for Calibrating Pretest Items.* (ETS RR–14-38), Princeton, NJ: Educational Testing Service, 2014. 12 p.

ANATCHKOVA, M. D.; ROSE, M.; WARE JR., J. E.; BJORNER, J. B. Development of an item bank and computer adaptive test for role functioning. *Qualityof Life Research***,** v. 21, n. 9, p. 1625-1637, 2012.

ANDERSSON, B.; BRANBERG, K.; WIBERG, M. Performing the Kernel Method of Test Equating with the Package kequate. *Journal of Statistical Software*, v.55, n. 6, p.1-25, 2013.

ANDRADE, D. F.; TAVARES, H. R.; VALLE, R da C. *Teoria da Resposta ao Item:* Conceitos e Aplicações. Caxambú-MG: 14 SINAPE, 2000.

ANDRIOLA, W. B. Descrição dos Principais Métodos para Detectar o Funcionamento Diferencial dos Itens (DIF). *Psicologia: Reflexão e Crítica*, v. 14, n. 3, p. 643-652, 2001.

_____. Estudo sobre o viés de itens em testes de rendimento: uma retrospectiva. *Estudos em Avaliação Educacional*, v. 17, n. 35, p. 115-134, 2006.

ARAÚJO, E. A. C.; ANDRADE, D. F; BORTOLOTTI, S. L. V. Teoria da Resposta ao Item. *Revista da Escola de Enfermagem*, v. 43, p. 1000-1008, 2009.

ARIEL, A.; VAN DER LINDEN, W. J.; VELDKAMP, B. P. A. Strategy for Optimizing Item-Pool Management. *Journal of Educational Measurement*, v. 43, n. 2, p. 85-96, 2006.

BAKER, F. B. *The Basics of Item Response Theory*. 2 ed. University of Wisconsin: ERIC, 2001.

BAN, J.C.; HANSON, B. A.; WANG, T.; YI, Q.; HARRIS, D. J. A Comparative Study of On-Line Pretest Item: Calibration/Scaling Methods in Computerized Adaptive Testing. *Journal of Educational Measurement***,** v. 38, n. 3, p. 191-212, 2001.

BAN, J.C.; HANSON, B. A.; YI, Q.; HARRIS, D. J. Data sparseness and on-line pretest item calibration-scaling methods in CAT. *Journal of Educational Measurement*, v. 39, n. 3, p. 207–218, 2002.

BARRADA J. R.; ABAD, F.J.; VELDKAMP, B. Comparison of methods for controlling maximum exposure rates in computerized adaptive testing. *Psicothema*, v. 21, n. 2, p. 313-20, 2009.

BARRADA, J. R.; MAZUELA, P.; OLEA, J. Maximum information stratification method for controlling item exposure in computerized adaptive testing. *Psicothema***,** v. 18, n. 1, p. 156-159, 2006.

BARRADA, J. R.; OLEA, J.; PONSODA, V. Methods for Restricting Maximum Exposure Rate in Computerized Adaptative Testing. *Methodology*, v. 3, n. 1, p. 14-23, 2007.

BARRADA, J. R.; OLEA, J.; PONSODA, V.; ABAD, F. J. Incorporating randomness to the Fisher information for improving item exposure control in CATS. *British Journal of Mathematical and Statistical Psychology*, v. 61, p. 493-513, 2008.

_____. Test overlap rate and item exposure rate as indicators of test security in CATs. In: WEISS, D. J. (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*, 2009.

_____. A method for the comparison of item selection rules in computerized adaptive testing. *Applied Psychological Measurement*, v. 34, n. 6, p. 438-452, 2010.

BATTAUZ, M. *equateIRT:* Direct, Chain and Average Equating Coefficients with Standard Errors Using IRT Methods. R package version 1.2, 2014.

BEKMAN, R. M. Aplicação dos blocos incompletos balanceados na teoria de resposta ao item. *Estudos em Avaliação Educacional*, n. 24, p. 119-138, 2001.

BELOV, D. I. *Detection of Large-Scale Item Preknowledge in Computerized Adaptive Testing via Kullback–Leibler Divergence.* Law School Admission Council Research Report 12-01, p. 1-25, 2012.

BELOV, D. I.; ARMSTRONG, R. D. Direct and Inverse Problems of Item Pool Design for Computerized Adaptive Testing. *Educational and Psychological Measurement*, v. 69, n. 4, p. 533-547, 2009.

BELOV, D. I.; ARMSTRONG, R. D.; WEISSMAN, A. A Monte Carlo approach for adaptive testing with content constraints. *Applied Psychological Measurement,* v. 32, n. 6, p. 431-446, 2008.

BERGER, M. P. F. D-optimal sequential sampling designs for item response theory models. *Journal of Educational Statistics*, v. 19, p. 43-56, 1994.

BERGSTROM, B. A.; GERSHON, R. C. Item Banking. In: IMPARA, J. C. (Ed.). *Licensure Testing: Purposes, Procedures, and Practices.* Lincoln, NE: Buros, 1995, Disponível em: http://digitalcommons.unl.edu/ buroslicensure/13. Acesso em: ago. 2014.

BERGSTROM, B.; GERSHON, R.; LUNZ, M. E. Computerized Adaptive Testing Exploring Examinee Response Time Using Hierarchical Linear Modeling. 26p. In: *Annual meeting National Council on Measurement in Education,* New Orleans, Louisiana, 1994.

BIRNBAUM, A. Some Latent Trait Models and Their Use in Infering an Examinee's Ability. In: LORD, F.M.; NOVICK, M. R**.** *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley, 1968. p. 395-479.

BJORNER, J. B. CHANG, C. H.; THISSEN, D.; REEVE, B.B. Developing tailored instruments: item banking and computerized adaptive assessment. *Quality of Life Research,* v. 16, p. 95-108, 2007.

BOCK, R. D.; LIEBERMAN, M. Fitting a response model for dichotomously scored items. *Psychometrika*, v. 35, n. 2, p. 179-197, 1970.

BOCK, R. D.; MURAKI, E.; PFEIFFENBERGER, W. Item Pool Maintenance in the Presence of Item Parameter Drift. *Journal of Educational Measurement*, v. 25, n. 4, p. 275-285, 1988.

BOCK, R.D.; AITKIN, M. Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika*, v. 46, p. 443-459, 1981.

BOCK, R.D.; GIBBONS, R.; MURAKI, E. Full-Information item factor analysis. *Applied Psychological Measurement*, v. 12, n. 3, p. 261-280, 1988.

BRASIL. *Guia de elaboração e revisão de itens*. v. 1. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira – INEP; 2010.

CELLA, D.; GERSHON, R.; LAI, J-S.; CHOI, S. The future of outcomes measurement: Item banking, tailored short forms, and computerized adaptive assessment. *Quality of Life Research*, v. 16, p. 133-141, 2007.

CERVANTES, V. H. *DFIT:* An R package for the Differential Functioning of Items and Tests framework. Instituto Colombiano para la Evaluacion de la Educacion, Bogota, Colombia, 2014.

CHALMERS, P. *mirtCAT:* Computerized Adaptive Testing with Multidimensional Item Response Theory. R package version 0.5, 2015.

_____. mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, v. 48, n. 6, p. 1-29, 2012.

CHANG, H. H. Psychometrics behind Computerized Adaptive Testing. *Psychometrika*, v. 80, n. 1, p. 1-20, 2015.

CHANG, H. H.; QIAN, J.; YING, Z. a-Stratified multistage computerized adaptive testing with b blocking. *Applied Psychological Measurement*, v. 25, p. 333-341, 2001.

CHANG, H. H.; YING, Z. A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, v. 20, n. 3, p. 213-229, 1996.

_____. a-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, v. 23, p. 211-222, 1999.

_____. Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. *Annals of Statistics*, v. 37, n. 3, p. 1466-1488, 2009.

CHANG, S-R.; PLAKE, B.S.; FERDOUS, A. A. Response Times for Correct and Incorrect Item Responses on Computerized Adaptive Tests. In: *Annual meeting of the American Educational Research Association*, Montréal, Canada, 2005.

CHANG, S-W.; ANSLEY, T. N. A Comparative Study of Item Exposure Control Methods in Computerized Adaptive Testing. *Journal of Educational Measurement*, v. 40, n. 1, p. 71-103, 2003.

CHANG, W.; CHENG, J.; ALLAIRE, J.; XIE, Y.; MCPHERSON, J. *shiny:* Web Application Framework for R. R package version 0.12.2., 2015.

CHANG, Y. C. I.; LU, H. Y. Online calibration via variable length computerized adaptive testing. *Psychometrika*, v. 75, n. 1, p. 140-157, 2010.

CHEN, S. A procedure for controlling general test overlap in computerized adaptive testing. *Applied Psychological Measurement*, v. 34, n. 6, p. 393-409, 2010.

CHEN, S. Y.; ANKENMANN, R. D.; CHANG, H. H. A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement*, v. 24, n. 3, p. 241-255, 2000.

CHEN, S-Y.; ANKENMANN, R. D.; SPRAY, J. A. The relationship between item exposure and test overlap in Computerized Adaptive Testing. *Journal of Educational Measurement*, v. 40, n. 2, p. 129-145, 2003.

CHEN, S-Y.; LEI, P-W. Controlling Item Exposure and Test Overlap in Computerized Adaptive Testing. *Applied Psychological Measurement*, v. 29, n. 3, p. 204–217, 2005.

_____. Investigating the relationship between item exposure and test overlap: Item sharing and item pooling. *British Journal of Mathematical and Statistical Psychology*, v. 63, p. 205-226, 2010.

CHEN, S-Y.; LEI, P-W.; CHEN, J-H.; LIU, T-C. General Test Overlap Control: Improved Algorithm for CAT and CCT. *Applied Psychological Measurement*, v. 38, n. 3, p. 229-244, 2014.

CHEN, S-Y.; LEI, P-W.; LIAO, W-H. Controlling item exposure and test overlap on the fly in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, v. 61, p. 471-492, 2008.

CHENG, P. E.; LIOU, M. Estimation of trait level in computerized adaptive testing. *Applied Psychological Measurement*, v. 24, n. 3, p. 257-265, 2000.

_____. Computerized adaptive testing using the nearest-neighbors criterion. *Applied Psychological Measurement*, v. 27, n. 3, p. 204-216, 2003.

CHOI, S. W.; GIBBONS, L. E.; CRANE, P. K. lordif: An R Package for Detecting Differential Item Functioning Using Iterative Hybrid Ordinal Logistic Regression/Item Response Theory and Monte Carlo Simulations. *Journal of Statistical Software*, v. 39, n. 8, p. 1-30, 2011.

CHOI, S. W.; KING, D. R. *MAT:* Multidimensional Adaptive Testing. R package version 2.2., 2014.

CHOI, S. W.; SWARTZ, R. J. Comparison of CAT Item Selection Criteria for Polytomous Items. *Applied Psychological Measurement*, v. 33, n. 6, p. 419-440, 2009.

CLARK, A. Review of Parameter Drift Methodology and Implications for Operational Testing. In: *National Conference of Bar Examiners*. p. 1-21, 2013. Disponível em: http://www.ncbex.org/assets/media_files/Research/2013Clark.pdf. Acesso em: agosto de 2014.

COHEN, R. J.; SWERDLIK, M. E.; STURMAN, E. D. *Testagem e Avaliação Psicológica:* Introdução a Testes e Medidas, 8. ed., Porto Alegre: AMGH Editora Ltda., 756 p., 2014.

CONCERTO. *Open-source online adaptive testing plataform*. The Psychometrics Centre.University of Cambridge, 2016. Disponível em: http://concertoplatform.com/. Acesso em: dez. 2016.

COSTA, D. R., KARINO, C. A., MOURA, F. A. S., ANDRADE, D. F. A comparison of three methods of item selection for computerized

adaptive testing. In: WEISS, D. J. (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*, 2009.

DAVEY, T. *A Guide to Computer Adaptive Testing (CAT) Systems.* Washington, DC: Council of Chief State School Officers, 2011, 31 p.

DAVEY, T.; PARSHALL, C. G. New algorithms for item selection and exposure control with computerized adaptive testing. In: *Annual meeting of the American Educational Research Association*, San Francisco, CA, 1995.

DAVEY, T.; POMMERICH, M.; THOMPSON, T. D. Pretesting alongside an Operational CAT. In: *Annual Meeting of the National Council on Measurement in Education*, Montreal, Quebec, Canada, 1999, 22 p.

DE AYALA, R. J. *The theory and practice of item response theory.* New York, NY: Guilford, 2009.

DEMARS, C. E. Detection of item parameter drift over multiple test administrations. *Applied Measurement in Education*, v. 17, n. 3, p. 265-300, 2004.

_____. *Item response theory.* Series in understanding statistics. Oxford University Press, Inc., 2010.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, v. 39, p.1-38, 1977.

DOEBLER, A. The Problem of Bias in Person Parameter Estimation in Adaptive Testing. *Applied Psychological Measurement*, v. 36, n. 4, p. 255-270, 2012.

DONOGHUE, J. R.; ISHAM, S. P. A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement*, v. 22, n. 1, p. 33-51, 1998.

DORANS, N. J.; KULICK, E. Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, v. 23, p. 355-368, 1986.

DRASGOW, F.; LEVINE, M. V.; WILLIAMS, E. A. Appropriateness measurement with polychotomous item response models and

standardized indices. *British Journal of Mathematical and Statistical Psychology*, v. 38, p. 67-86, 1985.

EDWARDS, M. C. An Introduction to Item Response Theory Using the Need for Cognition Scale. *Social and Personality Psychology Compass*, v. 3, n. 4, p. 507-529, 2009.

ETS. Educational Testing Service. *ETS Standards for Quality and Fairness*, 2014/2015. Disponível em: https://www.ets.org/s/about/pdf/standards.pdf. Acesso em: 15 jan. 2016.

FAN, Z.; WANG, C.; CHANG, H.; DOUGLAS, J. Utilizing response time distributions for item selection in CAT. *Journal of Educational and Behavioral Statistics*, v. 37, p. 655-670, 2012.

FETZER, M.; DAINIS, A.; LAMBERT, S.; MEADE, A. *Computer Adaptive Testing (CAT) in an Employment Context*. SHL PreVisor. White Paper, p. 1-14, 2011.

FINKELMAN, M. D.; KIM, W.; WEISSMAN, A; COOK, R. J. Cognitive Diagnostic Models and Computerized Adaptive Testing: Two New Item-Selection Methods That Incorporate Response Times. *Journal of Computerized Adaptive Testing*, v. 2, n. 4, p. 59-76, 2014.

FINKELMAN, M.; NERING, M. L.; ROUSSOS, L. A Conditional Exposure Control Method for Multidimensional Adaptive Testing. *Journal of Educational Measurement*, v. 46, n. 1, p. 84-103, 2009.

FLIEGE, H.; BECKER, J.; WALTER, O.B.; ROSE, M.; BJORNER, J.B.; KLAPP, B. F. Evaluation of a computer-adaptive test for the assessment of depression (D-CAT) in clinical application. *International Journal of Methods in Psychiatric Research*, v. 18, n. 1, p. 23-36, 2009.

FRICK, H.; STROBL, C.; LEISCH, F.; ZEILEIS, A. Flexible Rasch Mixture Models with Package psychomix. *Journal of Statistical Software*, v. 48, n. 7, p. 1-25, 2012.

GEORGIADOU, E.; TRIANTAFILLOU, E.; ECONOMIDES, A. A. A Review of Item Exposure Control Strategies for Computerized Adaptive Testing Developed from 1983 to 2005. *The Journal of Technology, Learning, and Assessment*, v. 5, n. 8, p. 1-38, 2007.

GLAS, C. A. W. Item Parameter Estimation and Item Fit Analysis. In: VAN DER LINDEN, W.J.; GLAS, C.A.W. (Eds.). *Elements of Adaptive*

*Testing*. Statistics for Social and Behavioral Sciences, New York: Springer, 2010. p. 269-288.

GLAS, C. A. W.; VAN DER LINDEN, W. J. Computerized adaptive testing with item cloning. *Applied Psychological Measurement*, v. 27, n. 4, p. 247-261, 2003.

GONZALEZ, J. SNSequate: Standard and Nonstandard Statistical Models and Methods for Test Equating. *Journal of Statistical Software*, v. 59, n. 7, p. 1-30, 2014.

GONZÁLEZ-BETANZOS, F.; ABAD, F. J.; BARRADA, J. R. Fixed item parameter calibration for assessing differential item functioning in computerized adaptive tests. *Psicológica*, v. 35, p. 331-359, 2014.

GREEN, B. F.; BOCK, R. D.; HUMPHREYS, L. G.; LINN, R. L.; RECKASE, M. D. Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, v. 21, n.4, p. 347-360, 1984.

GUO, F.; WANG, L. Online calibration and scale stability of a CAT program. In: *Annual meeting of the National Council on Measurement in Education*, Chicago, IL, 2003.

GUO, R. *Item parameter drift and online calibration.* Tese (Doctor of Philosophy), Psychology in the Graduate College, University of Illinois, Urbana-Champaign, 2016.

GWALTNEY, C. J.; SHIELDS, A. L.; SHIFFMAN, S. Equivalence of Electronic and Paper-and-Pencil Administration of Patient-Reported Outcome Measures: A Meta-Analytic Review. *Value in Health*, v. 11, n. 2, p. 322-333, 2008.

HALADYNA TM. *Developing and validating multiple choice test items.* 3 ed. Mahwah, NJ: Lawrence Erlbaum; 2004.

HAMBLETON, R. K.; SWAMINATHAN, H.; ROGERS, J. *Fundamentals of Item Response Theory*. Measurement Methods for the Social Science. Newbury Park, CA: SAGE Publications, 1991. 184 p.

HAMBLETON, R. K.; ZAAL, J. N.; PIETERS, J. P. M. Computerized Adaptive Testing: Theory, Applications, and Standards. In: HAMBLETON, R. K.; ZAAL, J. N. (Eds.). *Advances in Educational and*

*Psychological Testing: Theory and Applications.* New York: Springer, 1991, p. 341-366.

HAN, K. T.; GUO, F. *Potential Impact of Item Parameter Drift Due to Practice and Curriculum Change on Item Calibration in Computerized Adaptive Testing*. GMAC Research Reports, RR-11-02, Reston, Virginia, p. 1-10, 2011.

HARMES, J. C.; PARSHALL, C. G.; KROMREY, J. D. Recalibration of IRT item parameters in a CAT: Sparse data matrices and missing data treatments. In: *Annual Meeting of the National Council on Measurement in Education*, Chicago: IL, 2003.

HARNISS, M.; AMTMANN, D.; COOK, D.; JOHNSON, K. Considerations for Developing Interfaces for Collecting Patient-Reported Outcomes that allow the inclusion of Individuals with Disabilities. *Medical Care*, v. *45*(5Suppl 1), S48, 2007.

HART, D. L.; DEUTSCHER, D.; CRANE, P. K.; WANG, Y. C. Differential item functioning was negligible in an adaptive test of functional status for patients with knee impairments who spoke English or Hebrew. *Quality of Life Research*, v. 18, n. 8, p. 1067-1083, 2009.

HE, W.; DIAO, Q.; HAUSER, C. A Comparison of Four Item-Selection Methods for Severely Constrained CATs. *Educational and Psychological Measurement*, v. 74, n. 4, p. 677-696, 2014.

HOHENSINN, C. *pcIRT*: IRT models for polytomous and continuous item responses. R package version 0.2., 2015.

HOLLAND, P. W.; THAYER, D. T. Differential item performance and the Mantel-Haenszel procedure. In: WAINER, H.; BRAUN, H. (Eds.). *Test validity*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988. p. 129-145.

HOLLAND, P. W.; WAINER, H. *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1993.

HORNKE, L. F. Item Response Times in Computerized Adaptive Testing. *Psicológica*, v. 21, p. 175-189, 2000.

HUANG, Y-M.; LIN, Y-T.; CHENG, S-C. An adaptive testing system for supporting versatile educational assessment. *Computers and Education*, v. 52, n. 1, p. 53-67, 2009.

HUFF, K. L.; SIRECI, S. G. *Validity issues in computer-based testing*. American Institute of Certified Public Accountants- AICPA, Technical Report, 2000.

HUNG, M.; STUART, A.R.; HIGGINS, T.F.; SALTZMAN, C.L.; KUBIAK, E.N. Computerized Adaptive Testing Using the PROMIS Physical Function Item Bank Reduces Test Burden With Less Ceiling Effects Compared With the Short Musculoskeletal Function Assessment in Orthopaedic Trauma Patients. *Journal of Orthopaedic Trauma,* v. 28, n. 8, p. 439-443, 2014.

IACAT. *International Association of Computerized Adaptive Testing*, 2015. Disponível em: http://iacat.org/content/operational-cat-programs. Acesso em: ago. 2015.

INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira**.** *Guia de elaboração e revisão de itens.* BRASÍLIA-DF, 2010.

ITO, K.; SYKES, R. C. The effect of restricting ability distributions in the estimation of item difficulties: Implications for a CAT implementation. In: *Annual meeting of the National Council on Measurement in Education*, New Orleans, LA, 1994.

JETTE, A. M.; HALEY, S. M.; NI, P.; OLARSCH, S.; MOED, R. Creating a computer adaptive test version of the late-life function and disability instrument. *Journals of Gerontology* - Series A Biological Sciences and Medical Sciences, v. 63, n. 11, p. 1246-1256, 2008.

KALINOWSKI, K. E.; NATESAN, P.; HENSON, R. K. Stratified Item Selection and Exposure Control in Unidimensional Adaptive Testing in the Presence of Two-Dimensional Data. *Applied Psychological Measurement,* v. 38, n. 7, p. 563-576, 2014.

KAMEI-HANNAN, C. Examining the Accessibility of a Computerized Adapted Test Using Assistive Technology. *Journal of Visual Impairment & Blindness*,v. 102, n. 5, p. 261-271, 2008.

KAYA, Z.; TAN, S. New trends of measurement and assessment in distance education. *Turkish Online Journal of Distance Education,* v. 15, n. 1, p. 206-217, 2014.

KELDERMAN, H. Item bias detection using loglinear IRT. *Psychometrika,* v. 54, p. 681-697, 1990.

KIEFER, T.; ROBITZSCH, A.; WU, M. *TAM*: Test Analysis Modules. R package version 1.13-0, 2015.

KIM, S. A Comparative Study of IRT Fixed Parameter Calibration Methods. *Journal of Educational Measurement*, v. 43, n. 4, p. 355-381, 2006.

KINGSBURY, G. G. Adaptive item calibration: A process for estimating item parameters within a computerized adaptive test. In: WEISS, D. J. (Ed.). *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*. Disponível em: www.psych.umn.edu/psylabs/CATCentral. Acesso em: jul. 2014.

KINGSBURY, G. G.; ZARA, A. R. Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, v. 2, n. 4, p. 359-375, 1989.

KOPEC, J. A.; BADII, M.; MCKENNA, M.; LIMA, V. D.; SAYRE, E. C. DVORAK, M. Computerized adaptive testing in back pain – Validation of the CAT-5D-QOL. *Spine*, v. 33, n. 12, p. 1384-1390, 2008.

KRASS, I. A.; WILLIAMS, B. Calibrating CAT Pools and Online Pretest Items Using Nonparametric and Adjusted Marginal Maximum Likelihood Methods. In: *Annual Meeting of the National Council on Measurement in Education*, Chicago, IL, 2003.

LAI, J. S.; CELLA, D.; CHOI, S. et al. How Item Banks and Their Application Can Influence Measurement Practice in Rehabilitation Medicine: A PROMIS Fatigue Item Bank Example. *Archives of Physical Medicine and Rehabilitation*, v. 92, n. 10, p. S20-S27, 2011.

LEI, P. W.; CHEN, S. Y.; YU, L. Comparing Methods of Assessing Differential Item Functioning in a Computerized Adaptive Testing Environment. *Journal of Educational Measurement*, v. 43, n. 3, p. 245-264, 2006.

LEROUX, A. J.; LOPEZ, M.; HEMBRY, I.; DODD, B. G. A Comparison of Exposure Control Procedures in CATs Using the 3PL Model. *Educational and Psychological Measurement*, v. 73, n. 5, p. 857-874, 2013.

LI, X. An Investigation of the Item Parameter Drift in the Examination for the Certificate of Proficiency in English (ECPE). *Spaan Fellow*

*Working Papers in Second or Foreign Language Assessment*, v. 6, p. 1-28, 2008.

LORD, F. M. Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika*, v. 39, p. 247-264, 1974.

_____. *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1980.

LU, H.Y. Application of Optimal Designs to Item Calibration. *PLOS ONE*, v. 9, n. 9, p. 1-8, 2014.

LU, Y.; HAMBLETON, R. K. Statistics for detecting disclosed items in a CAT environment**.** *Metodología de las ciencias del comportamiento*, v. 5, n. 2, p. 225-242, 2004.

LUECHT, R. M.; DE CHAMPLAIN, A.; NUNGESTER, R. J. Maintaining Content Validity in Computerized Adaptive Testing. *Advances in Health Sciences Education*, v. 3, n. 1, p. 29-41, 1998.

MADAUS, G. F. The Distortion of Teaching and Testing: High-Stakes Testing and Instruction. *Peabody Journal of Education*, v. 65, n. 3, p. 29-46, 1988.

MAGIS, D. A note on the item information function of the four-parameter logistic model. *Applied Psychological Measurement*, v. 37, p. 304-315, 2013.

MAGIS, D.; BARRADA, J. R. *Open-source CAT software:* R packages and Concerto. SOQOL-NL, Utrecht, 2014. Disponível em: http://www.isoqol.nl/sites/default/files/Symposia%20bestanden/David%20Magis%20Open%20source%20CAT%20software.pdf. Acessoem: ago. 2014.

MAGIS, D.; BELAND, S.; RAICHE, G. *difR:* Collection of methods to detect dichotomous differential item functioning (DIF). R package version 4.6, 2015.

MAGIS, D.; RAICHE, G. Random Generation of Response Patterns under Computerized Adaptive Testing with the R Package catR. *Journal of Statistical Software*, v. 48, n. 8, p. 1-31, 2012.

MAIR, P.; HATZINGER, R.; MAIER M. J. *eRm:* Extended Rasch Modeling. 0.15-5., 2015.

MAKRANSKY, G., GLAS, G. A. W. An automatic online calibration design in adaptive testing. *Journal of Applied Testing Technology*, v. 11, n. 1, p. 1-29, 2010.

_____. Modeling differential item functioning with group-specific item parameters: A computerized adaptive testing application. *Measurement: Journal of the International Measurement Confederation*, v. 46, n. 9, p. 3228-3237, 2013.

MANSEIRA; P. R. P.; MISAGHI, M. Proposta de ferramenta para uso abrangente de Testes Adaptativos Computadorizados na Educação a Distância. In: *Congresso Brasileiro de Engenharia de Produção*, Ponta Grossa, PR, Brasil, 2013.

MANTEL, N., HAENSZEL, W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, v. 22, p. 719-748, 1959.

MASTERS, J. S.; MUCKLE, T. J.; BONTEMPO, B. Comparing Methods to Recalibrate Drifting Items in Computerized Adaptive Testing. In: *Conference American Educational Research Association*, San Diego, CA, p. 1-28, 2009.

MATTEUCCI, M.; VELDKAMP, B. P. On the use of MCMC computerized adaptive testing with empirical prior information to improve efficiency. *Statistical Methods and Applications*, v. 22, n. 2, p. 243-267, 2013.

MAZZA, A.; PUNZO, A.; MCGUIRE, B. KernSmoothIRT: An R Package for Kernel Smoothing in Item Response Theory. *Journal of Statistical Software*, v. 58, n. 6, p. 1-34, 2014.

MCBRIDE, J. R.; MARTIN, J. T. Reliability and validity of adaptive ability tests in a military setting. In WEISS, D. J. (Ed.), *New horizons in testing:* Latent trait test theory and computerized adaptive testing.New York, NY: Academic Press, 1983. p. 224-236.

MCCLARTY, K. L.; SPERLING, R. A.; DODD, B. G. A variant of the progressive-restricted item exposure control procedure in computerized adaptive testing systems based on the 3PL and partial credit models. In: *Annual meeting of the American Educational Research Association*, San Francisco, CA, 2006.

MCCOY, K. M. *The impact of item parameter drift on examinee ability measures in a computer adaptive environment*. Tese (Doctor of Philosophy), University of Illinois, Chicago, IL, 2010.

MCDONOUGH, C. M.; TIAN, F.; NI, P. et al. Development of the computer-adaptive version of the late-life function and disability instrument. *Journals of Gerontology* - Series A Biological Sciences and Medical Sciences, v. 67, n. 12, p. 1427-1438, 2012.

MCLEOD, L. D.; LEWIS, C. Detecting item memorization in the CAT environment. *Applied Psychological Measurement*, v. 23, n. 2, p. 147-160, 1999.

MCLEOD, L. D.; LEWIS, C.; THISSEN, D. A Bayesian method for the detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement*, v. 27, n. 2, p. 121-137, 2003.

MCLEOD, L. D.; SCHNIPKE, D.L. detecting items that have been memorized in the Computerized Adaptive Testing Environment. In: *Annual Meeting of the National Council on Measurement in Education*, Montreal, Quebec, Canada, 1999. 23p.

MEIJER, R. R. Outlier detection in high-stakes certification testing. *Journal of Educational Measurement*, v. 39, n. 3, p. 219-233, 2002.

MEIJER, R. R.; VAN KRIMPEN-STOOP, E. M.L.A. Detecting Person Misfit in Adaptive Testing. VAN DER LINDEN, W.J.; GLAS, C.A.W. (Eds.). *Elements of Adaptive Testing*, Statistics for Social and Behavioral Sciences. New York: Springer, 2010, p. 315-329.

METERKO, M.; MARFEO, E.E.; McDonough, C.M. et al. Work Disability Functional Assessment Battery: Feasibility and Psychometric Properties. *Archives of Physical Medicine and Rehabilitation*, v. 96, n. 6, p. 1028-1035, 2015.

MILLER, T. R. Practical considerations for conducting studies of differential item functioning (DIF) in a CAT environment. In: *Annual meeting of the American Educational Research Association*, San Francisco, CA, 1992.

MILLS, C. N.; STOCKING, M. L. *Practical issues in large-scale high-stakes computerized adaptive testing*. (Research Report RR-95-23), Princeton, NJ: Educational Testing Service, 1995. 30 p.

MISLEVY, R. L. Bayes modal estimation in item response theory. *Psychometrika*, v. 51, p. 177-195, 1986.

MOREIRA JUNIOR, F. de J. *Sistemática para a implantação de testes adaptativos informatizados baseados na teoria da resposta ao item*. Tese (Doutorado em Engenharia de Produção), Universidade Federal de Santa Catarina, Florianópolis, 2011. 334 p.

MURPHY, D. L.; DODD, B. G.; VAUGHN, B. K. A comparison of item selection techniques for testlets. *Applied Psychological Measurement*, v. 34, n. 6, p. 424-437, 2010.

NANDAKUMAR, R.; ROUSSOS, L. Evaluation of the CATSIB DIF procedure in a pretest setting. *Journal of Educational and Behavioral Statistics*, v. 29, n. 2, p. 177-199, 2004.

NERING, M. L., DAVEY, T., THOMPSON, T. A hybrid method for controlling item exposure in computerized adaptive testing. In: *Annual meeting of the Psychometric Society*, Urbana, IL, 1998.

NUNES, C. H. S. S.; PRIMI, R. Impacto do tamanho da amostra na calibração de itens e estimativa de traços latentes por teoria de resposta ao item. *Avaliação Psicológica*, v. 4, n. 2, p. 141-153, 2005.

NUNES, C.H.S.S.; SPENASSATO, D.; OLIVEIRA, C.M.; BORNIA, A.C.; PRIMI, R. Testes Adaptativos Computadorizados - CAT. In: SILVA, M.C.R.; BATHOLOMEU, D.; VENDRAMINI, C.M.M.; MONTIEL, J.M. (Eds.). *Aplicações de métodos estatísticos avançados à avaliação psicológica e educacional*. 1 ed. São Paulo: Vetor, 2015, p. 37-76.

NYDICK, S. W. *catIrt:* An R Package for Simulating IRT-Based Computerized Adaptive Tests. R package version 0.5-0., 2014.

OWEN, R. J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, v. 70, p. 351-356, 1975.

OZYURT, H.; OZYURT, O.; BAKI, A.; GUVEN, B. An Application of Individualized Assessment in Educational Hypermedia: Design of Computerized Adaptive Testing System and its Integration Into UZWEBMAT. *Procedia - Social and Behavioral Sciences*, v. 46, p. 3191-3196, 2012.

PASQUALI, L. Validade dos Testes Psicológicos: Será Possível Reencontrar o Caminho? *Psicologia: Teoria e Pesquisa*, v. 23, n. especial, p. 99-107, 2007.

_____. *Psicometria:* teoria dos testes na psicologia e na educação. 5. ed. Petrópolis, RJ: Vozes, 2013.

PILKONIS, P. A.; LAN, Y.; DODDS, N. E.; JOHNSTON, K. L.; MAIHOEFER, C. C.; LAWRENCE, S. M. Validation of the depression item bank from the Patient-Reported Outcomes Measurement Information System (PROMIS®) in a three-month observational study. *Journal of Psychiatric Research*, v. 56, p. 112-119, 2014.

PIROMSOMBAT, C. *Differential Item Functioning in Computerized Adaptive Testing: Can CAT Self-Adjust Enough?*. 192 p. Tese (Doctor of Philosophy). University of Minnesota, 2014.

PITON GONÇALVES, J. *Desafios e perspectivas da implementação computacional de testes adaptativos multidimensionais para avaliações educacionais*. 153 p. Tese (Doutorado em Ciências - Ciências de Computação e Matemática Computacional), Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2013.

PITON GONÇALVES, J.; ALUÍSIO, S. M. Multidimensional Computer Adaptive test with educational purposes: principles and methods. *Ensaio: Avaliação e Políticas Públicas em Educação*, v. 23, n. 87, p. 389-414, 2015.

POMMERICH, M.; SEGALL, D. O. Calibrating CAT Pools and Online Pretest Items Using Marginal Maximum Likelihood Methods. In: *Annual Meeting of the National Council on Measurement in Education*, Chicago, IL, 2003. p. 1-33.

PREINERSTORFER, D. *mRm:* An R package for conditional maximumlikelihood estimation in mixed Rasch models. R package version 1.1.5, 2013.

PRIMI R.; MUNIZ, M.; NUNES, C.H.S.S. *Definições contemporâneas de validade de testes psicológicos.* Programa de Pós Graduação em Psicologia, Universidade São Francisco.18 p. Disponível em: http://unipe.br/blog/psicologia/wp-content/uploads/2012/10/ validade_de_testes_psicologicos.pdf. Acesso em: jan. 2016.

QIAN, H.; STANIEWSKA, D.; RECKASE, M.; WOO, A. Using Response Time to Detect Item Preknowledge in Computer-Based Licensure Examinations. *Educational Measurement: Issues and Practice*, v. 35, n. 1, p. 38-47, 2016.

R CORE TEAM. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. Disponível em: http://www.R-project.org/.

RAJU, N. S. The área between two item characteristic curves. *Psychometrika*, v. 53, p. 495-502, 1988.

RAJU, N. S.; VAN DER LINDEN, W. J.; FLEER, P. F. IRT-based internal measures of diferential functioning of items and tests. *Applied Psychological Measurement*, v.19, p. 353-368, 1995.

REEVE, B.B.; HAYS, R.D.; BJORNER, J.B. et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*, v. 45, p. S22–S31, 2007.

REIF, M. *mcIRT:* IRT models for multiple choice items. R package version 0.41, 2014.

REVELLE, W. *psych*: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, Version 1.5.1, 2015.

REVUELTA, J.; PONSODA, V. A Comparison of Item Exposure Control Methods in Computerized Adaptive Testing. *Journal of Educational Measurement*, v. 35, n. 4, p. 311-327, 1998.

RILEY, B. B.; CARLE, A. C. Comparison of two Bayesian methods to detect mode effects between paper-based and computerized adaptive assessments: A preliminary Monte Carlo study. *BMC Medical Research Methodology*, v. 12, p. 1-13, 2012.

RISK, N. M. *The Impact of Item Parameter Drift in Computer Adaptive Testing (CAT).* Tese (Doctor of Philosophy), Educational Psychology in the Graduate College, University of Illinois, Chicago, 2015.

RIZOPOULOS, D. ltm*:* An R package for Latent Variable Modelling and Item Response Theory Analyses, *Journal of Statistical Software*, v. 17, n. 5, p. 1-25, 2006.

ROBITZSCH, A. *sirt*: Supplementary Item Response Theory Models. R package version 1.8-9, 2015.

ROGERS, J.; SWAMINATHAN, H. A logistic regression procedure for detecting item bias. In: *Annual meeting of the American Educational Research Association*, San Francisco, 1989.

RUDICK, M. M.; YAM, W. H.; SIMMS, L. J. Comparing countdown- and IRT-based approaches to computerized adaptive personality testing. *Psychological Assessment,* v. 25, n. 3, p. 769-779, 2013.

SCHNIPKE, D. L.; GREEN, B. F. A comparison of item selection routines in linear and adaptive tests. *Journal of Educational Measurement,* v. 32, n. 3, p. 227-242, 1995.

SEGALL, D. O. Calibrating CAT Pools and Online Pretest Items Using MCMC Methods. In: *Annual meeting of the National Council on Measurement in Education, Chicago, IL*, p. 1-9, 2003.

_____. A sharing item response theory model for computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, v. 29, n. 4, p. 439-460, 2004.

_____. Computerized Adaptive Testing. *Encyclopedia of Social Measurement,* v. 1, p. 429-438, 2005.

SHEALY, R.T.; STOUT, W. F. An Item Response Theory Model for Test Bias and Differential Test Functioning. In: HOLLAND, P. W.; WAINER, H. (Eds.). *Differential Item Functioning*. New Jersey: Lawrence Erlbaum Associates, Inc., p. 197-239, 1993.

SIRECI, S. G., WAINER, H., THISSEN, D. On the reliability of testlet-based tests. *Journal of Educational Measurement*, v. 28, p. 237-247, 1991.

SMITS, N.; CUIJPERS, P.; VAN STRATEN, A. Applying computerized adaptive testing to the CES-D scale: A simulation study. *Psychiatry Research*, v. 188, n. 1, p. 147-155, 2011.

SOLÓRZANO; R. W. High Stakes Testing: Issues, Implications, and Remedies for English Language Learners. *Review of Educational Research*, v. 78, n. 2, p. 260-329, 2008.

SPENASSATO, D., BORNIA, A. C., TEZZA, R. Computerized Adaptive Testing: A Review of Research and Technical Characteristics. *IEEE Latin America Transactions*, v. 13, n. 12, p. 3890-3898, 2015.

SQUIRES, P. *An Item Bank Approach to Testing*. Concept paper prepared by Applied Skills & Knowledge, LLC. Morristown, NJ, 2003. Disponível em: http://www.appliedskills.com/white_papers.html. Acesso em: ago. 2014.

STOCKING, M. L. *Some considerations in maintaining adaptive test item pools.* (Tech Rep. No. ERIC ED391814), Princeton, NJ: Educational Testing Service, 1988a.

_____. *Scale drift in on-line calibration.* (Research Report 88- 28 -ONR), Princeton, NJ: Educational Testing Service, 1988b. 134p.

_____. Specifying optimum examinees for item parameter estimation in item response theory. *Psychometrika*, v. 55, n. 3, p. 461-475, 1990.

_____. *Three Practical Issues for Modern Adaptive Testing Item Pools.* (Report No ETS-RR-94-5), Princeton, NJ: Educational Testing Service, 1994, 45 p.

STOCKING, M.L.; LEWIS, C.L. Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, v. 23, p. 57-75, 1998.

STOCKING M. L.; LEWIS, C. Methods of Controlling the Exposure of Items in CAT. In VAN DER LINDEN, W.J.; GLAS, C.A.W. (Eds.). *Computerized Adaptive Testing:* Theory and Practice. Netherlands: Kluwer Academic Publishers, 2000. p. 163-182.

STONE, E; LAITUSIS, C. C.; COOK, L. L. Increasing the accessibility of assessments through technology. In: DRASGOW, F. (Ed.). *Technology and Testing:* Improving Educational and Psychological Measurement. New York: Routledge, 2015, p. 217-234.

STRAETMANS, G. J. J. M.; EGGEN, T. J. H. M. Comparison of Test Administration Procedures for Placement Decisions in a Mathematics Course. *Educational Research and Evaluation,* v. 4, n. 3, p. 259-275, 1998.

SWAMINATHAN, H.; GIFFORD, J.A. Bayesian estimation in the three-parameter logistic model. *Psychometrika*, v. 51, p. 589-601, 1986.

SWANSON, D. B.; FEATHERMAN, C.; CASE, S. M.; LUECHT, R. M.; NUNGESTER, R. J. Relation of response latency to test design, examinee proficiency, and item difficulty in computer-based test administration. In: *Annual Meeting of the National Council on Measurement in Education*, Chicago, IL, 1997.

SYMPSON, J. B.; HETTER, R. D. Controlling item-exposure rates in computerized adaptive testing. In: *Annual meeting of the military testing association*, Navy personnel research and development center, San Diego, CA, p. 973-977, 1985.

TEZZA, R.; BORNIA, A. C.; ANDRADE, D. F. Measuring web usability using item response theory: Principles, features and opportunities. *Interacting with Computers*, v. 23, n. 2, p. 167-175, 2011.

THISSEN, D.; MOONEY, J. Loglinear item response models, with applications to data from social surveys. *Sociological Methodology*, v. 19, p. 299-330, 1989.

THISSEN, D.; REEVE, B. B.; BJORNER, J. B.; CHANG, C. H. Methodological issues for building item banks and computerized adaptive scales. *Quality of Life Research*, v. 16, p. 109-119, 2007.

THISSEN, D.; STEINBERG, L.; WAINER, H. Use of item response theory in the study of group differences in trace lines. In: WAINER, H.; BRAUN, H. (Eds.). *Test validity*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988. p. 147-169.

_____. Detection of diferential item functioning using the parameters of item response models. In: HOLLAND, P.W.; WAINER, H. (Eds.). *Diferential item functioning*, Hillsdale, NJ: Erlbaum, 1993, p. 67-113.

THOMAS, N.; GAN, N. Generating multiple imputations for matrix sampling data analyzed with item response models. *Journal of Educational and Behavioral Statistics*, v. 22, n. 4, p. 425-445, 1997.

THOMAS, T. J. Item-presentation controls for multidimensional item pools in computerized adaptive testing. *Behavior Research Methods, Instruments & Computers*, v. 22, n. 2, p. 247-252, 1990.

THOMPSON, N. A.; WEISS, D. J. A Framework for the Development of Computerized Adaptive Tests. *Practical Assessment, Research & Evaluation*, v. 16, n. 1, p. 1-9, 2011.

TIAN, J-Q.; MIAO, D-M.; ZHU, X.; GONG, J-J. An Introduction to the Computerized Adaptive Testing. *US-China Education Review*, v. 4, n. 1, p. 72-81, 2007.

URBINA, S. *Fundamentos da Testagem Psicológica.* Porto Alegre: Artmed, 2007. 320 p.

URRY, V. W. *A Monte Carlo investigation of logistic test models.* Tese (Doctoral dissertation), Purdue University, West Lafayette, 1970.

VAN DER ARK, L. A. New Developments in Mokken Scale Analysis in R. *Journal of Statistical Software*, v. 48, n. 5, p. 1-27, 2012.

VAN DER LINDEN, W. J. Bayesian item selection criteria for adaptive testing. *Psychometrika*, v. 63, n. 2, p. 201-216, 1998.

_____. Empirical initialization of the trait estimator in adaptive testing. *Applied Psychological Measurement*, v. 23, n. 1, p. 21-29, 1999.

_____. Some alternatives to Sympson-Hetter item-exposure control in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, v. 28, p. 249-265, 2003.

_____. A Comparison of Item-Selection Methods for Adaptive Tests with Content Constraints. *Journal of Educational Measurement*, v. 42, n. 3, p. 283-302, 2005.

_____. A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, v. 31, p. 181-204, 2006.

_____. Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, v. 33, n. 1, p. 5-20, 2008.

_____. Constrained Adaptive Testing with Shadow Tests. In: VAN DER LINDEN, W.J.; GLAS, C.A.W. (Eds.). *Elements of Adaptive Testing, Statistics for Social and Behavioral Sciences*, 2010. p. 31-55.

VAN DER LINDEN, W. J.; GUO, F. Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, v. 73, p. 365-384, 2008.

VAN DER LINDEN, W. J.; PASHLEY, P. J. Item Selection and Ability Estimation in Adaptive Testing. VAN DER LINDEN, W.J.; GLAS,

C.A.W (Eds.). *Elements of Adaptive Testing*, Statistics for Social and Behavioral Sciences, New York: Springer, 2010. p. 3-30.

VAN DER LINDEN, W. J.; REESE, L. M. A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, v. 22, n. 3, p. 259-270, 1998.

VAN DER LINDEN, W. J.; REN, H. Optimal Bayesian Adaptive Design for test-item calibration. *Psychometrika*, v. 80, n. 2, p. 263-288, 2015.

VAN DER LINDEN, W. J.; SCRAMS, D. J.; SCHNIPKE, D. L. Using response-time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement*, v. 23, n. 3, p. 195-210, 1999.

VAN DER LINDEN, W. J.; VAN KRIMPEN-STOOP, E. M. L. A. Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika*, v. 68, n. 2, p. 251-265, 2003.

VAN DER LINDEN, W. J.; VELDKAMP, B. P. Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics*, v. 29, n. 3, p. 273-291, 2004.

_____. Conditional item-exposure control in adaptive testing using item-ineligibility probabilities. *Journal of Educational and Behavioral Statistics*, v. 32, p. 398-418, 2007.

VAN KRIMPEN-STOOP, E.M.L.A.; MEIJER, R.R. The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement*, v. 23, n. 4, p. 327-345, 1999.

_____. CUSUM-based person-fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics*, v. 26, n. 2, p. 199-217, 2001.

VEERKAMP, W. J. J.; BERGER, M. P. F. Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics*, v. 22, p. 203-226, 1997.

VEERKAMP, W. J. J.; GLAS, C. A. W. Detection of known items in adaptive testing with a statistical quality control method. *Journal of Educational and Behavioral Statistics*, v. 25, n. 4, p. 373-389, 2000.

VELDKAMP, B. P.; MATTEUCCI, M. Bayesian computerized adaptive testing. *Ensaio*, v. 21, n. 78, p. 57-82, 2013.

VELDKAMP, B. P.; VAN DER LINDEN, W. J. Designing item pools for computerized adaptive testing. In: VAN DER LINDEN, W.J.; GLAS, C.A.W. (Eds.). *Computerized adaptive testing:* Theory and practice, Boston: Kluwer Academic, 2000. p. 149-162.

VISPOEL, W. P. Psychometric characteristics of computer-adaptive and self-adaptive vocabulary tests: The role of Answer Feedback and Test Anxiety. *Journal of Educational Measurement,* v. 35, n. 2, p. 155-167, 1998.

WAINER, H. CATs: whit her and whence. *Psicologica,* v. 21, p. 121-133, 2000.

WAINER, H.; MISLEVY, R.J. Item response theory, calibration, and estimation. In: WAINER, H. (Ed.). *Computerized Adaptive Testing:* A Primer. Mahwah: Lawrence Erlbaum Associates, 2000.

WALKER, C. M.; BERETVAS, S. N.; ACKERMAN, T. An Examination of Conditioning Variables Used in Computer Adaptive Testing for DIF Analyses. *Applied Measurement in Education,* v. 14, n. 1, p. 3-16, 2001.

WALKER, J.; BÖHNKE, J. R.; CERNY, T.; STRASSER, F. Development of symptom assessments utilising item response theory and computer-adaptive testing-A practical method based on a systematic review. *Critical Reviews in Oncology/Hematology,* v. 73, n.1, p.47-67, 2010.

WANG, T. Essentially unbiased EAP estimates in computerized adaptive testing. In: A*nnual meeting of the American Educational Research Association,* Chicago, IL, 1997.

WANG, T.; HANSON, B. A.; LAU, C. M. A. Reducing bias in CAT trait estimation: A comparison of approaches. *Applied Psychological Measurement,* v. 23, n. 3, p. 263-278, 1999.

WANG, T.; KOLEN, M. J. Evaluating Comparability in Computerized Adaptive Testing: Issues, Criteria and an Example. *Journal of Educational Measurement,* v. 38, n. 1, p. 19-49, 2001.

WANG, T.; VISPOEL, W. P. Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement,* v. 35, n. 2, p. 109-135, 1998.

WARD, A. W.; MURRAY-WARD, M. An NCME Instructional Module: Guidelines for the Development of Item Banks. *Educational Measurement: Issues and Practice*, v. 13, n. 1, p. 34-39, 1994.

WARM, T. A. Weighted likelihood estimation of ability in item response theory. *Psychometrika*, v. 54, n. 3, p. 427-450, 1989.

WAUTERS, K.; DESMET, P.; VAN DEN NOORTGATE, W. Adaptive item-based learning environments based on the item response theory: possibilities and challenges. *Journal of Computer Assisted Learning*, v. 26, p. 549-562, 2010.

WAY, W. D. Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice*, n. 17, p. 17-27, 1998.

_____. *Practical Questions in Introducing Computerized Adaptive Testing for K-12 Assessments*. (Research Report 05-03), Pearson Educational Measurement, p. 1-17, 2006.

WEISSMAN, A. A feedback control strategy for enhancing item selection efficiency in computerized adaptive testing. *Applied Psychological Measurement*, v. 30, n. 2, p. 84-99, 2006.

WELLS, C. S.; HAMBLETON, R. K.; KIRKPATRICK, R.; MENG, Y. An Examination of Two Procedures for Identifying Consequential Item Parameter Drift. *Applied Measurement in Education*, v. 27, n. 3, p. 214-231, 2014.

WELLS, C. S.; SUBKOVIAK, M. J.; SERLIN, R. C. The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement*, v. 26, n. 1, p. 77-87, 2002.

WISE, S. L. Overview of practical issues in a CAT program. In: *Annual meeting of the National Council on Measurement in Education*, Chicago, IL, 1997.

_____. The Utility of Adaptive Testing in Addressing the Problem of Unmotivated Examinees. *Journal of Computerized Adaptive Testing*, v. 2, n. 1, p. 1-17, 2014.

WISE, S. L.; KINGSBURY, G. G. Practical Issues in Developing and Maintaining a Computerized Adaptive Testing Program. *Psicológica*, v. 21, p. 135-155, 2000.

WISE, S. L.; KONG, X. Response time effort: A new measure of examinee motivation in computer-cased tests. *Applied Measurement in Education*, v. 16, p. 163-183, 2005.

WISE, S. L.; MA, L. Setting Response Time Thresholds for a CAT Item Pool: The Normative Threshold Method. In: *Annual meeting of the National Council on Measurement in Education*, Canada, 2012.

WOLLACK, J. A.; COHEN, A. S.; WELLS, C. S. A method for maintaining scale stability in the presence of test speededness. *Journal of Educational Measurement*, v. 40, n. 4, p. 307-330, 2003.

YEN, W. M. Effect of local dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, v. 8, p. 125–145, 1984.

YI, Q.; CHANG, H.H. a-Stratified CAT design with content blocking. *British Journal of Mathematical and Statistical Psychology*, v. 56, n. 2, p. 359-78, 2003.

YI, Q.; WANG, T.; BAN, J. C. Effects of scale transformation and test-termination rule on the precision of ability estimation in computerized adaptive testing. *Journal of Educational Measurement*, v. 38, n. 3, p. 267-292, 2001.

YI, Q.; ZHANG, J.; CHANG, H. Identifying practical indices for enhancing item pool security. In: *Annual meeting of the National Council on Measurement in Education*, Montreal, Canada, 2005.

ZHANG, J. A Sequential Procedure for Detecting Compromised Items in the Item Pool of a CAT System. *Applied Psychological Measurement*, v. 38, n. 2, p. 87-104, 2014.

ZHANG, J.; LI, J. Monitoring Items in Real Time to Enhance CAT Security. *Journal of Educational Measurement*, v. 53, n. 2, p. 131-151, 2016.

ZHAO, Y; HAMBLETON, R. *Software for IRT Analyses:* Descriptions and Features. 2009. Disponível em: http://hbanaszak.mjr.uw.edu.pl/TempTxt/IRTSoftwareReview.pdf. Acesso em: set. 2015.

ZHENG, Y. *New methods of online calibration for item bank replenishment.* Tese (Doctor of Philosophy in Educational Psychology), University of Illinois, Urbana-Champaign, 2014.

ZHENG, Y.; CHANG, C. H.; CHANG, H. H. Content-balancing strategy in bifactor computerized adaptive patient-reported outcome measurement. *Quality of Life Research*, v. 22, n. 3, p. 491-499, 2013.

ZHU, R. *Implementation of optimal design for item calibration in computerized adaptive testing (CAT).* Tese (Doctor of Educational measurement), University of Illinois, Urbana- Champaign, 2006.

ZHU, R., YU, F., LIU, S. Statistical indexes for monitoring item behavior under computer adaptive testing environment. In: *Annual meeting of the American Educational Research Association*, New Orleans, 2002.

ZIMOWSKI, M.; MURAKI, E.; MISLEVY, R.; BOCK, D. *Software BILOG-MG V3.0.* Scientific Software International, Inc., 2003.

ZITNY, P.; HALAMA, P.; JELÍNEK, M.; KVĚTON, P. Validity of cognitive ability tests - comparison of computerized adaptive testing with paper and pencil and computer-based forms of administrations. *Studia Psychologica,* v. 54, n. 3, p. 181-194, 2012.

ZWICK, R. The Investigation of Differential Item Functioning in Adaptive Tests. In: VAN DER LINDEN, W.J.; GLAS, C.A.W. (Eds.). *Elements of Adaptive Testing*, Statistics for Social and Behavioral Sciences, New York: Springer, 2010. p. 331-352.

ZWICK, R.; THAYER, D. T. Application of an Empirical Bayes Enhancement of Mantel-Haenszel Differential Item Functioning Analysis to a Computerized Adaptive Test. *Applied Psychological Measurement,* v. 26, n. 1, p. 57-76, 2002.

ZWICK, R.; THAYER, D.T.; WINGERSKY, M. *A simulation study of methods for assessing differential item functioning in computer-adaptive tests*. (ETS Research Report No. 93-11), Princeton, N J: Educational Testing Service, 1993.

_____. A simulation study of methods for assessing differential item functioning in computerized adaptive tests. *Applied Psychological Measurement,* v. 18, n. 2, p. 121-140, 1994a.

_____. *DIF analysis for pretest items in computer adaptive testing.* (ETS Research Report 94-33), Princeton, NJ: Educational Testing Service, 1994b.